SVM Feature Selection for Classification of SPECT Images of Alzheimer's Disease using Spatial Information

Jonathan Stoeckel
Siemens Medical Solutions USA
Computer Aided Diagnosis
Malvern, PA 19355
jonathan.stoeckel@siemens.com

Glenn Fung
Siemens Medical Solutions USA
Computer Aided Diagnosis
Malvern, PA 19355
glenn.fung@siemens.com

Abstract

Alzheimer's disease is the most frequent type of dementia for elderly patients. Due to aging populations the occurrence of this disease will increase in the next years. Early diagnosis is crucial to be able to develop more powerful treatments. Brain perfusion changes can be a marker for Alzheimer's disease. In this article we study the use of SPECT perfusion imaging for the diagnosis of Alzheimer's disease differentiating between images from healthy subjects and images from Alzheimer's disease patients. Our classification approach is based on a linear programming formulation similar to the 1-norm support vector machines. In contrast with other linear hyperplane-based methods that perform simultaneous feature selection and classification, our proposed formulation incorporates proximity information about the features and generates a classifier that does not just select the most relevant voxels but the most relevant "areas" for classification resulting in more robust classifiers that are better suitable for interpretation.

This approach is compared with the classical Fisher linear discriminant (FLD) classifier as well as with statistical parametric mapping (SPM).

We tested our method on data from four European institutions. Our method achieved sensitivity of 84.4% at 90.9% specificity, this is considerable better the human experts. Our method also outperformed the FLD and SPM techniques. We conclude that our approach has the potential to be a useful help for clinicians.

1 Introduction

Alzheimer's disease (AD) is the most frequent type of dementia for elderly patients. Due to aging populations its occurrence will still increase. Even though no definitive cure has been found for this disease, reliable diagnosis is useful for excluding other dementias, choosing the right treatment and for the development of new treatments.

AD is diagnosed using the criteria from the National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) [1]. In practice the main tool for evaluating patients are neuro-psychologic tests, that test abilities like memory and language. The Mini Mental State Examination (MMSE) is the most widely used of these tests [2].

Brain images can also provide some helpful indication of AD. Magnetic resonance imaging (MRI) is used to study possible anatomical changes of the brain [3]. Images showing the local perfusion (amount of blood flow) of the brain can be used for the diagnosis of AD because the perfusion pattern is affected by the disease. In this article we will look into the use of cerebral perfusion imaging acquired by single photon emitting computer tomography (SPECT) using technetium-99m hexamethylpropylene amine oxime (HMPAO) as the tracer. SPECT imaging is a largely accepted clinical modality for AD diagnosis. Even though the perfusion pattern and its evolution is not the same for all patients some hypo-perfusion patterns seem to be typical for the disease. There are three main regions mentioned in literature attained by hypo-perfusion[4], 1. the temporo-parietal region, 2. the posterior cingulate gyri and precunei, and 3. the medial temporal lobe. The first region is known as the predominant pattern for AD, however this region was not found for early AD [5]. The second region is probably more specific and more frequent in early AD [6]. Previous pathological studies have suggested that the third region is the first affected by the disease [7], however in practice it is only observed in more advanced stages of the disease [6].

There is not one single perfusion pattern that differentiates AD patients form healthy subjects. Thus it might be useful to have tools that could assist physicians in this difficult task. In this article we will present a method that does not need any explicit knowledge about the perfusion pattern of AD patients.

Some approaches for a computer aided diagnosis (CAD) system for the analysis of SPECT images for AD can be found in literature. The first family is based on the analysis of regions of interest. The mean values for these regions are analyzed using some discriminant functions (see e.g. [8][9]).

The second approach is statistical parametric mapping (SPM) and its numerous variants. Statistical parametric mapping is widely used in the neuro-sciences. Its framework was first developed for the analysis of SPECT and PET studies, but is now mainly used for the analysis of functional MRI data. It was not developed specifically to study a single image, but for comparing groups of images. One can use it for diagnostics by comparing the image under study to a group of normal images.

Statistical parametric mapping consists of doing a voxelwise statistical test, in our case a t-test, comparing the values of the image under study to the mean values of the group of normal images. Subsequently the significant voxels are inferred by using random field theory (see e.g. [10] for a full description of SPM). A largely used freely available implementation called SPM99 [11] has been developed and is used in this article as comparison to our approach.

In this article we will propose another approach using as less a-priori information about the pathology as possible, by obtaining it implicitly from image databases. Another important aspect is that our approach is global. that all the information in the image can be used at once in contrast to more local approaches, e.g mono-variate methods like SPM. A multi-variate approach generally increases sensitivity at the price of loosing regional specificity (e.g. depicting local hypo-perfusion regions). However in the approach presented in this paper compared to our earlier work [12] we use feature selection while trying to add spatial constraints to the classification.

The following section first discusses the pre-processing of the data, next we describe our proposed mathematical programming formulation. Unlike the traditional SVM-like formulations, spatial information about the feature (voxels) locations is incorporated into the optimization problem. This leads to feature selection where the classifier depends on regions in the brain instead of isolated non-connected voxels. In section 3 we present the data we used for our experiments. It consists of real brain SPECT images obtained from four different institutions. The results on the data are presented in section 4 and discussed in section 5.

1.1 Notation

We now describe the notation used in this paper. The notation $A \in R^{m \times n}$ will signify a real $m \times n$ matrix. For such a matrix, A' will denote the transpose of A and A_i will denote the i-th row of A. All vectors will be column vectors. For $x \in R^n$, $\|x\|_p$ denotes the p-norm, $p = 1, 2, \infty$. A vector of ones in a real space of arbitrary dimension will be denoted by e. Thus, for $e \in R^m$ and $y \in R^m$, $e'y$ is the sum of the components of y. A vector of zeros in a real space of arbitrary dimension will be denoted by 0. A *separating hyperplane*, with respect to two given point sets $\mathcal{A}$ and $\mathcal{B}$, is a plane that attempts to separate R^n into two halfspaces such that each open halfspace contains points mostly of $\mathcal{A}$ or $\mathcal{B}$.

2 Methods

2.1 Spatial Normalization

In the classifier based approach we need the assumption that the same position in the volume coordinate system within different volumes corresponds to the same anatomical position. This makes it possible to do meaningful voxelwise comparisons between images. However this assumption is not met by the images without pre-processing: First of all, the subject which is being imaged, is not always positioned at the same position in the reference frame of the imaging device. This reference frame defines where e.g. the brain is positioned in the image. Secondly the anatomy does not always have the same shape and size between different subjects. For example, the size and shape of the skull can already be largely different between subjects. This means that we need to spatially register the volumes. In our application we do not have detailed knowledge of the anatomy of our subjects as only HMPAO-SPECT images of the subjects were available. These images are so-called functional images. They only depict the regional blood flow of the subject. The regional cerebral blood flow provides us of course with some gross information about the anatomy, but only based on the fact that there is a relationship between the blood flow, and the underlying anatomy. Understanding this characteristic of HMPAO SPECT images is fundamental for the choice of the registration method.

Because of the limited anatomical information available in the volumes we chose to estimate affine transformations between the volumes and not use transformations with a larger number of degrees of freedom. We used the correlation ratio as the similarity measure [13] that we minimized using Powell optimization [14]. To obtain a more robust result we used the following procedure. First of all, we registered all volumes to a single volume, then we calculated a mean volume. This mean volume was first put on the midsagittal plane by registering it with a flipped version (see [15]). Subsequently it was made to be symmetrical by taking the mean of itself with a flipped version. Finally all volumes were matched to this volume.

2.2 Intensity Normalization

HMPAO SPECT imaging generates volumes that only give a relative measure of the blood flow. The blood flow measure is relative to the blood flow in other regions of the brain. Direct comparison, of the voxel intensities, between images, even different acquisitions of the same subject, is thus not possible without normalization of the intensities.

For all the experiments, we normalize the intensities by applying an affine transformation to the intensities. The transformation parameters are estimated on the training set of each experiment such that the intensities for each voxel position have zero mean and standard deviation of one for all the training subjects. We choose this very common data normalization since it provides numerical stability to the algorithms involved.

2.3 Classification

Because the hypo-perfusion pattern for early AD is not very well defined we choose to develop a method where we do not use any explicit knowledge about the typical perfusion patterns. We use implicit knowledge about the perfusion patterns by using a database of images of AD patients and normal subjects. To separate the images we use a classifier using the voxel intensities as features and this database to train the classifier. Using the voxel intensities as features makes it possible not to introduce any particular knowledge about the exact location of the hypo-perfusion area(s). Thus by using a database of images and the voxel intensities we circumvent the problem of the exact definition of the typical perfusion pattern for early AD. In general the number of images available in the training databases is significantly smaller (< 100) than the number of voxels (> 1000). Thus the number of features (voxels) is much larger than the number of samples (training images). The number of samples is considered to be small if it is about the same or smaller than the number of dimensions. In this case we speak of almost empty spaces, the small sample size problem or the so called curse of dimensionality. In classical pattern recognition it is believed that no good generalization could be obtained for these cases when using the whole feature space [16]. Generalization is the capacity of a classifier to rightly classify a sample never seen before. In order to improve generalization of our final classifier, minimal feature dependency (small amount of features) of the classifier is desired.

2.3.1 The Linear Support Vector Machine

We consider the problem, depicted in Figure 1, of classifying m points in the n-dimensional real space R^n, represented by the $m \times n$ matrix A, according to membership of each point A_i in the class $A+$ or $A-$ as specified by a given $m \times m$ diagonal matrix D with plus ones or minus ones along its diagonal. For this problem the standard support vector machine with a linear kernel [16] is given by the following quadratic program with parameter $\nu > 0$:

$$\begin{array}{ll} \min\limits_{(w,\gamma,y)\in R^{n+1+m}} & \nu e'y + \frac{1}{2}w'w \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & y \geq 0. \end{array} \tag{1}$$

Here, the plane $x'w = \gamma + 1$ bounds the class $A+$ points, while the plane $x'w = \gamma - 1$ bounds the class $A-$ points as follows:

$$\begin{array}{llll} A_i w & \geq & \gamma + 1, & \text{for} \quad D_{ii} = 1 \\ A_i w & \leq & \gamma - 1, & \text{for} \quad D_{ii} = -1. \end{array} \tag{2}$$

The linear separating surface is the plane $x'w = \gamma$ midway between the bounding planes (2). The quadratic term in (1) maximizes the distance or "margin" between the bounding planes. Maximizing the margin enhances the generalization capability of a support vector machine [16]. In order to make use of a faster linear programming based approach, instead of the standard quadratic programming formulation (1), we reformulate (1) by replacing the 2-norm by a 1-norm as follows [17]:

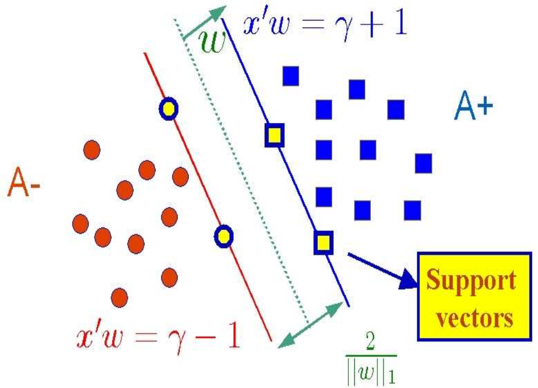

Figure 1. The approximately bounding planes of equation (2) with a soft (i.e. with some error) margin $\frac{2}{\|w\|_1}$, and the plane $x'w = \gamma$ approximately separating $A+$ from $A-$ are represented by the red, green and blue lines. In this case, the support vectors are the points that lie on the bounding planes.

$$\begin{array}{ll} \min\limits_{(w,\gamma,y)\in R^{n+1+m}} & \nu e'y + \|w\|_1 = \nu\sum\limits_{i=1}^{m} y_i + \sum\limits_{j=1}^{n} |w_j| \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & y \geq 0. \end{array} \tag{3}$$

This SVM$\| \cdot \|_1$ reformulation in effect maximizes the margin, the distance between the two bounding planes of Figure

1, using a different norm, the ∞-norm, and results with a margin in terms of the 1-norm, $\frac{2}{\|w\|_1}$, instead of $\frac{2}{\|w\|_2}$ [18]. The mathematical program (3) is easily converted to a linear program as follows:

$$\begin{array}{rl}\displaystyle\min_{(w,\gamma,y,v)\in R^{n+1+m+n}} & \nu e'y + e'v = \nu\sum_{i=1}^{m} y_i + \sum_{j=1}^{n} v_j \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & v \geq w \geq -v \\ & y \geq 0,\end{array} \tag{4}$$

Empirical evidence [17] indicates that the 1-norm formulation has the advantage of generating very sparse solutions. This results in the normal w to the separating plane $x'w = \gamma$ having many zero components, which implies that many input space features do not play a role in determining the linear classifier. This makes this approach suitable for feature selection in classification problems. We note that in addition to the conventional interpretation of smaller ν as emphasizing a larger margin between the bounding planes (2), a smaller ν here also results in a sparse solution. The "right" value of ν is determined by a tuning procedure where the performance is adjusted to the desired compromise between the classification performance and the sparseness of the solution. Next, we will revisit some regularization theory results that would motivate the SVM-like formulation we are proposing in this paper.

2.4 Regularization Theory and SVMs

Let $f : \Re^n \rightarrow \Re$ with $f(x) = w'x - \gamma$ the our prediction or classification function. Then, Formulation (4) and Support Vector Machine (SVM) formulations in general can be seen as a particular case of regularization networks [19] where the functional $R_{reg}[f] = R_{emp} + \lambda G(Pf)$ that is often referred as the regularized risk, is minimized. $R_{reg}[f]$ is equal to the empirical risk functional $R_{emp}[f]$ plus a regularization term $G(Pf)$ that is usually defined as $\|Pf\|^2$. $\lambda = \frac{1}{\nu}$ is the regularization parameter and P is a called the regularization operator. P maps the the classifier function f into some dot product space [20]. For example, in the case of SVMs, the type of regularization and the class of functions that form the basis for the prediction function are intimately related. The SVM algorithm is equivalent to minimizing $R_{reg}[f]$ on the family of functions $f(x) = \sum_i \alpha_i k(x_i, x) + b$ provided that the kernel k is chosen as a Green's function of $P * P$ [20]. For example, in Formulation (4) the regularization term is $G(Pf) = \|w\|_1$. and $K(x_i, x_j) = x_i' x_j$ (the linear kernel). Our proposed formulation also proposed to minimize the regularized risk $R_{reg}[f]$ but for a very specific linear regularization operator P that encodes prior information (in the form of spatial information) about the classification task at hand.

2.4.1 The Contiguous Linear SVM (CSVM)

There are two drawbacks related to standard SVM formulations, especially when they are applied to imaging classification problems. The first drawback is related to the fact that little or no spatial information about the imaging problem is incorporated into the optimization problem to solve, discarding very valuable information that could lead to better and more robust classifiers. In the case of imaging problems where the features are related to voxel/pixel intensities a relation can be predefined among the voxels using spatial information or previous knowledge about the problem. The second drawback is related to the interpretability of the results. In several applications a feature selection scheme is implemented not only to get sparse models but also to determine which of the input features are relevant for the classification task, leading to insights about the problem in question. For example in the problem that we are addressing in this article it is easier to interpret a final classifier depending on contiguous voxels defining regions than a subset of independent voxels with no apparent connection among them. Our goal in this paper is to incorporate spatial information about every voxel into the optimization problem in a manner that the final obtained hyperplane classifier depends on regions or clusters of features rather than on isolated voxels. Let's consider a similarity function r that defines binary relations among any two features (f_i, f_j) of any given training datapoint. Let R be a matrix such that:

$$R_{ij} = r(f_i, f_j) \in \{0, 1\}, i, j \in \{1, \ldots, n\}$$

We define now, $\hat{R} = R - I_{n\times n}$, $\hat{R}$ is the symmetric adjacency matrix of an undirected graph representing the relation among the features according to the relation function r. R is a pseudo-adjacency matrix of a graph where every node has a self-loop. For most problems in real life R is based on local relations and therefore it is a very sparse matrix (see e.g. Figure 2). The function r could be defined in a more general way, where instead of a binary relations it can be a similarity function or any other kind of function encoding extra information about the features or the datapoints in the training set.

In our specific case we choose the relation r to be defined by a $3 \times 3 \times 3$ mask defining the 26-closest neighbors of each voxel. Note that this very local simple mask allows to encode the sense of contiguity among voxels in a global sense across the whole volume. This mask size was chosen because it provided excellent results while maintaining the sparsity of the relation r A very simple but effective way to incorporate this extra information about the features into the 1-norm SVM formulation (4) is to use the relationship matrix R as a regularization operator and then minimize the the regularized risk:

$$R_{reg}[f] = R_{emp} + \frac{1}{\nu}\left\|R^{-1}w\right\|_1 \tag{5}$$

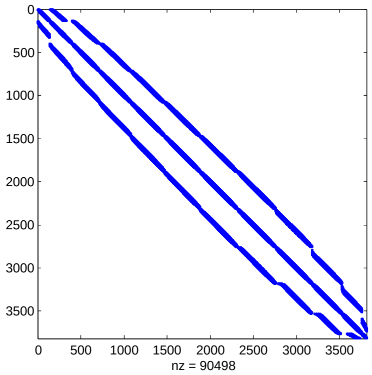

Figure 2. The sparse adjacency matrix R for the mask defining the 26-closest neighbors of each voxel.

This can be formulated as the following linear programming problem:

$$\begin{array}{rl} \min\limits_{(w,\gamma,y,v)\in R^{n+1+m+n}} & \nu e'y + e'v = \nu\sum\limits_{i=1}^{m} y_i + \sum\limits_{j=1}^{n} v_j \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & Rv \geq w \geq -Rv \\ & y \geq 0, \end{array} \tag{6}$$

At a solution of problem (4), v is the absolute value $|w|$ of w. This fact follows from the constraints $v \geq w \geq -v$ which imply that $v_i \geq |w_i|,\ i = 1\ldots,n$. Hence at optimality, $v = |w|$, otherwise the objective function can be strictly decreased without changing any variable except v. In this new formulation (4) we have at optimality that $Pv = |w|$, this is:

$$|w_i| = \sum_{j=1}^{n} R_{ij}vj = \sum_{\{j|ri,j=1\}} R_{ij}vj \tag{7}$$

In other words this means that the magnitude of the weight w_i of the related feature i, not only depends on itself but it also depends on all the features j that are related to i according to the relation function r. Moreover R can be interpreted as a covariance matrix such that the prior over the vector of weights w is given by $P(w) =\propto \exp(\left\|R^{-1}w\right\|_1)$.

3 Materials

3.1 Subjects

The images we used for our experiments were taken from a concurrent study investigating the use of SPECT as a diagnostic tool for the early onset of AD. A detailed description of this data can be found in [21]. Subjects of four different centers, Edinburgh (Scotland), Nice (France), Genoa (Italy), and Cologne (Germany) were included for this study. In total 158 subjects participated, including 99 patients with AD, 28 patients suffering from depression (not used in this article), and 31 healthy volunteers. An example of this data is seen in figure 3. Confirmation of Alzheimer's disease was obtained by clinical follow-up. There was no statistically significant age difference between the AD patients and the healthy subjects. For technical acquisition related reasons images of 7 AD subjects had to be excluded.

3.1.1 Pre-processing

Applying the registration procedure as described above results in images of 128 by 128 by 89 voxels, with a voxelsize of 1.71 mm by 1.71 mm by 1.88 mm for all four centers. The SPECT images have an effective resolution of about 7 mm full width at half maximum (FWHM). Therefore we can subsequently subsample the images a factor of two in each dimension by taking the average value over the subsampled areas without loosing much information. We only use the voxel intensities for the voxels in the part of the brain that has been imaged for all subjects. Applying this procedure results in 3816 features per subject available for classification/feature selection.

3.1.2 Experts

All real images were rated in four categories (very probable, probably, probably not and very unlikely to have AD) by sixteen European expert nuclear medicine physicians. The possible ratings were as follows: very probably Alzheimer's disease, probably Alzheimer's disease, probably not Alzheimer's disease and very unlikely Alzheimer's disease. To be able to compare the data from the experts with that of the automatic methods, we considered the first two ratings as positive and the other two as negative.

4 Experiments

In all of our experiments we divided the data into two disjoint training and testing sets. The idea is to tune the parameters in our model only using data from the training set, once the final model is fixed, it is tested in the unseen testing set. We used leave-one-out cross validation to tune the model parameter ν of the contiguous SVM. Performance of our Contiguous SVM algorithm, in terms of generalization ability, is compared with a Fisher's Linear Discriminant (FLD) classifier as previously presented in [12]. The FLD algorithm used here is based on the FLD mathematical programming formulation introduced by Mika et al ([22]).

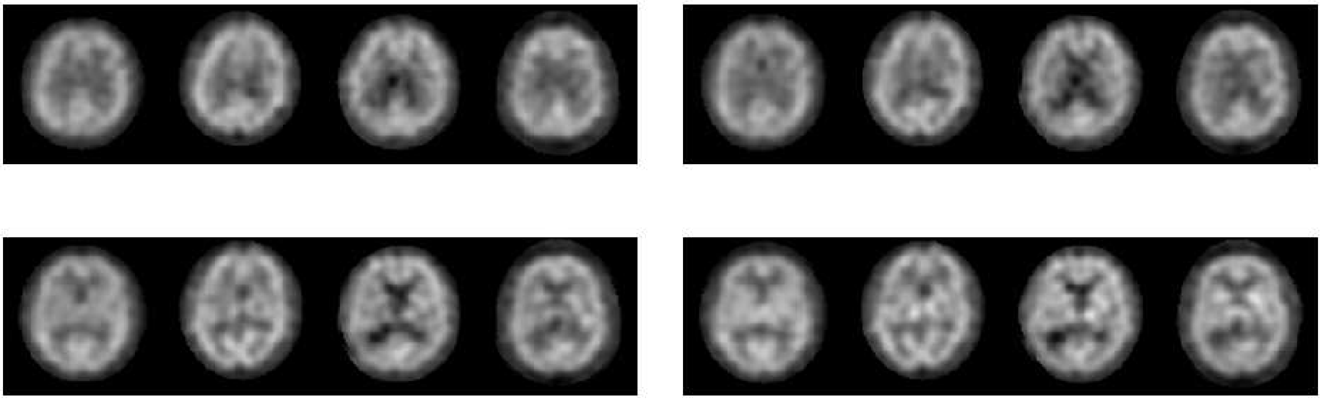

Figure 3. Examples of four volumes from Cologne after intensity and spatial normalization. In each column the first two small images show two normal subjects, the last two images show slices of AD subjects. The sets of slices are ordered from left to right and from top to bottom. Strong hypo-perfusion can be seen for the first AD patient, whereas the hypo-perfusion is more subtle for the second patient.

For solving all the optimization problems involved in this paper we used the widely used commercial solver CPLEX 6.5 [23]. Next, we outline the results of our comparative testing. Two set of experiments were performed:

1. We randomly divided the 123 cases into 90 training examples and 33 testing examples, the goal of this experiment is to approximately measure the generalization capability of our proposed classifier.

2. In order to test the generalization performance of our approach across institutions, we divided the data into two different subsets according to the institution where they were collected. The training set consists of 68 cases coming from Genoa (34 cases) and Cologne (34 cases) and the testing set consists of 55 cases coming from Edinburgh (28 cases) and Nice (27 cases).

The first experiment resulted in a selection of 253 features grouped in 7 connected areas. Figure 4 shows part of the selected features (a subset that can easily be visualized in 2D). Most selected groups of features are in the ventricles. This is consistent with the general atrophy of the brain observed in Alzheimer's disease patients which enlarges the ventricles relative to the other parts of the brain. This result shows the potential of the proposed approach at selecting meaningful grouped features which can be interpreted more easily than traditional feature selection approaches. The experts had an average sensitivity of 56.6% and a specificity of 82.4% for all 123 cases. In the SPM approach we use SPM at a significance level of 0.1 at the cluster level. We consider each image where some significant clusters were found to be a positive result, this leads to a sensitivity of 55.9% and a specificity of 77.4% for SPM. Our classification approach as shown in tables 1 and 2 outperforms both the experts and the SPM approach. Results in table 2 show that even if the performance decreases on the training set due to differences in the way the images were aqcuired at the different institutions the contiguous SVM approach still shows good generalization capabilities.

Table 1. Results for the first experiment for 90 training cases and 33 testing cases randomly sampled among the different institutions. The training results are based on leave-one out.

	CSVM Sensitivity Specificity	FLD Sensitivity Specificity
Training	86.7% **80.0**%	**88.7%** 65.0%
Testing	**84.4**% **90.9**%	82.0% 87.5%

5 Conclusion

Based on the experiments described in this article we conclude that our automatic approach to the classification of images performs at least as well as human observers. In general our contiguous support vector machine is more sen-

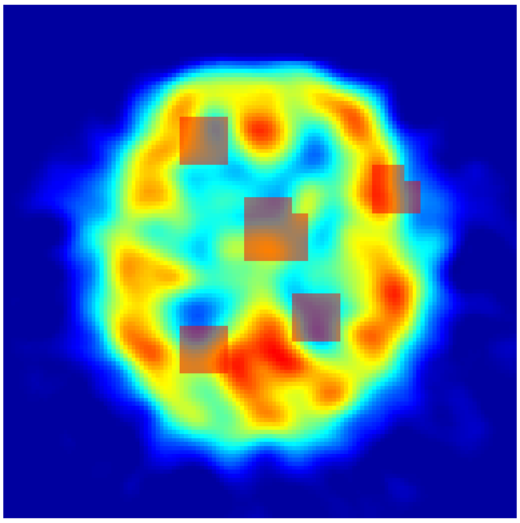

Figure 4. A single axial image showing the regions picked by the algorithm overlayed on an image of an Alzheimer's disease patient SPECT image.

sitive and more specific. One would need more data, especially of control subjects to be able to state that automatic methods always significantly outperform human observers in clinical practice. We have shown that classification of images using the voxel values as features outperforms the local SPM approach. We have shown that classification without using any specific knowledge related to the pathology is possible. The approach we propose in this article gives only a global decision based on a specific image. However only providing global information might not be sufficient for clinicians. Therefore we proposed a method that might do useful feature selection which might provide useful information to the clinician, at least at the group level. A trained classifier represents the group of images it was trained on, it does not show which areas where discriminative for any specific single image. Further research should focus on how to obtain subject specific local information while still retaining the advantage of a global approach. For future work one might want to try the presented approach for differential diagnosis (other dementias versus Alzheimer's disease) which might be an even more important clinical issue. ROC analysis of the classifier as well as of the experts will be useful to better compare performances. This will also provide means to handle the differences in operating points for the different experts (e.g. some experts are more specific while others are more sensitive). Also an interesting future direction would be to extend the Contiguous SVM formulation, where a relation among datapoints is considered instead of a relation among the features. This approach can potentially be used for a general semi-supervised SVM approach where only some of the labels for the training data are available.

Table 2. Results for the second experiment. The classifier was trained on the data from Genoa (34 cases) and Cologne (34 cases), and tested on the data from Edinburgh (28 cases) and Nice (27 cases). The training results are based on leave-one out.

	CSVM Sensitivity Specificity	FLD Sensitivity Specificity
Training	**86.2**% **68.0**%	84.6% 62.5%
Testing	**72.5**% 93.0%	45.0% **100.0%**

6 Acknowledgements

We are indebted to Prof. Nicholas Ayache and Prof. Jacques Darcourt for their contributions to this work.

References

[1] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E.M. Stadlan. Mental and clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of the Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–944, July 1984.

[2] M.F. Folstein, S.E. Folstein, and P.R. McHugh. "Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, November 1975.

[3] K.M. Gosche, J.A. Mortimer, C.D. Smith, W.R. Markesbery, and D.A. Snowdon. Hippocampal volume as an index of Alzheimer neuropathology: findings from the Nun Study. *Neurology*, 58(10):1476–1482, May 2002.

[4] I. Goethals, C. van de Wiele, D. Slosman, and R. Dierckx. Brain SPET perfusion in early Alzheimer's disease: where to look? *European Journal of Nuclear Medicine*, 29(8):975–978, August 2002.

[5] W.A. Van Gool, G.J. Walstra, S. Teunisse, F.M. Van der Zant, H.C. Weinstein, and E.A. Van Royen.

Diagnosing Alzheimer's disease in elderly, mildly demented patients: the impact of routine single photon emission computed tomography. *Journal of Neurology*, 242(6):401–405, June 1995.

[6] D. Kogure, H. Matsuda, T. Ohnishi, T. Asada, M. Uno, T. Kunihiro, S. Nakano, and M. Takasaki. Longitudinal evaluation of early Alzheimer's disease using brain perfusion SPECT. *Journal of Nuclear Medicine*, 41(7):1155–1162, July 2000.

[7] H. Braak and E. Braak. Diagnostic criteria for neuropathologic assessment of Alzheimer's disease. *Neurobiology and Aging*, 18(4):S85–S88, July 1997.

[8] M.R. Dawson, A. Dobbs, H.R. Hooper, A.J. McEwan, J. Triscott, and J. Cooney. Artificial neural networks that use single-photon emission tomography to identify patients with probable Alzheimer's disease. *European Journal of Nuclear Medicine*, 21(12):1303–1311, December 1994.

[9] D. Hamilton, D. O'Mahony, J. Coffey, J. Murphy, N. O'Hare, P. Freyne, B. Walsh, and D. Coakley. Classification of mild Alzheimer's disease by artificial neural network analysis of SPET data. *Nuclear Medicine Communications*, 18(9):805–810, September 1997.

[10] R.S.J. Frackowiak, K.J. Friston, C.D. Frith, and R. Dolan. *Human Brain Function*. Academic Press, 1997.

[11] J. Ashburner, K. Friston, A. Holmes, and J.-B. Poline. Statistical Parametric Mapping, SPM'99. The Welcome Department of Cognitive Neurology. Institute of Neurology, University College London, 1999. Freely available at *http://www.fil.ion.ucl.ac.uk/spm*.

[12] J. Stoeckel, Malandain G., O. Migneco, P.M. Koulibaly, Robert P., N. Ayache, and J. Darcourt. Classification of SPECT images of normal subjects versus images of Alzheimer's disease patients. In W.J. Niessen and M.A. Viergever, editors, *4th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI'01)*, volume 2208 of *LNCS*, pages 666–674, Utrecht, The Netherlands, October 2001.

[13] A. Roche, G. Malandain, X. Pennec, and N. Ayache. The Correlation Ratio as a New Similarity Metric for Multimodal Image Registration. In W. M. Wells, A. C. F. Colchester, and S. Delp, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI'98)*, volume 1496 of *Lecture Notes in Computer Science*, pages 1115–1124, Boston, USA, October 1998.

[14] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes. The Art of Scientific Computing*. Cambridge University Press, 2nd edition, 1997.

[15] S. Prima, S. Ourselin, and N. Ayache. Computation of the mid-sagittal plane in 3D brain images. *IEEE Transaction on Medical Imaging*, 21(2):122–138, February 2002.

[16] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[17] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Machine Learning Proceedings of the Fifteenth International Conference(ICML '98)*, pages 82–90, San Francisco, California, 1998. Morgan Kaufmann. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps.

[18] O. L. Mangasarian. Arbitrary-norm separating plane. *Operations Research Letters*, 24:15–23, 1999. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/97-07r.ps.

[19] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

[20] A. Smola, P. L. Bartlett, B. Schölkopf, and J. Schuurmann (editors). *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 2000.

[21] D. Soonawala, T. Amin, K.P. Ebmeier, J.D. Steele, N.J. Dougall, J. Best, O. Migneco, F. Nobili, and K. Scheidhauer. Statistical parametric mapping of (99m)Tc-HMPAO-SPECT images for the diagnosis of Alzheimer's disease: normalizing to cerebellar tracer uptake. *Neuroimage*, 17(3):1193–1202, November 2002.

[22] Sebastian Mika, Gunnar Rätsch, and Klaus-Robert Müller. A mathematical programming approach to the kernel fisher algorithm. In *NIPS*, pages 591–597, 2000.

[23] ILOG CPLEX Division, 889 Alder Avenue, Incline Village, Nevada. *CPLEX Optimizer*, 2004. `http://www.cplex.com/`.