3D - Patch Based Feature Extraction for Alzheimer's Disease Diagnosis via FDG-PET Analysis

Anant Srivastava[a], Shibani Singh[a], Jie Zhang[a], Liang Mi[a], Dhruman Goradia[b], Kewei Chen[b], Eric Reiman[b], Yalin Wang[a,*], for the Alzheimer's Disease Neuroimaging Initiative

[a] *School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA*

[b] *Banner Alzheimer's Institute, Phoenix, AZ, USA*

[*] *To be submitted to Medical Image Analysis.*

Abstract: 249 words
Title: 116 characters
Pages: 42
Figures: 8
Tables: 1

Please address correspondence to:

Dr. Yalin Wang
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
P.O. Box 878809
Tempe, AZ 85287 USA

Phone: (480) 965-6871
Fax: (480) 965-2751
E-mail: ylwang@asu.edu

Data used in preparation of this article were obtained from the phase two of the Alzheimer's Disease Neuroimaging Initiative (ADNI2) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

Alzheimer's disease (AD), is a chronic neurodegenerative disease that usually starts slowly and gets worse over time. It is the cause of 60% to 70% of cases of dementia. There is growing interest in identifying brain image biomarkers that help evaluate AD risk presymptomatically. High-dimensional non-linear pattern classification methods have been applied to structural magnetic resonance images (MRI's) and used to discriminate between clinical groups in Alzheimer's progression. Using Fluorodeoxyglucose (FDG) positron emission tomography (PET) as the preferred imaging modality, this paper develops a 3D patch based machine learning model and uses it to perform six binary classification experiments across different (AD) diagnostic categories. Specifically, from the PET image scans features were extracted and learned using probablistic PCA by taking overlapping patches in and around the cerebral cortex and using them as features. Using AdaBoost as the preferred choice of classifier both methods try to utilize ^{18}F-FDG PET as a biological marker in the early diagnosis of Alzheimer's . Additional we investigate the involvement of rich demographic features (Apoe$E3$, Apoe$E4$ and Functional Activities Questionnaires (FAQ)) in classification. The experimental results on Alzheimer's Disease Neuroimaging initiative (ADNI) dataset demonstrate the effectiveness of both the proposed systems. The paper reports a very high accuracy when comparing AD to Cognitively Normal and concludes that for least group seperation in disease progession the approach needs to be adhoc and extremely fractional. The use of ^{18}F-FDG PET may offer a new sensitive biomarker and enrich the brain imaging analysis toolset for studying the diagnosis and prognosis of AD.
Keywords: Probablistic PCA, Max-pooling, Ada-boost, Alzheimer's Disease, ADNI2.

1. Introduction

Alzheimer's disease (AD) is a chronic neurodegenerative disease in which amyloid plaques and neurofibrillary tangles accumulate in the brain. The most common early symptom is the difficulty remembering recent events (short-term memory loss). As the disease advances, patients may lack motivation, have problems with self-care, and show behavioral abnormalities or even withdraw from family and society (Burns and Iliffe). AD has a typical

pattern of progression, with changes in the brain that correspond to the types and severity of symptoms. Disease progression has commonly been divided into five main categories in the ADNI2 phase of Alzheimer's Disease Neuroimaging initiative(ADNI) (Weiner et al., 2013): Cognitively Normal (CN), Significant Memory Concern(SMC), Early Mild Cognitive Impairment (EMCI), Impairment (LMCI) and Alzheimer's Disease (AD), defined clinically based on behavioral and cognitive assessment. (SMC) are self-report significant memory concern from the patient we choose to exclude it from our study.

There has been a shift with a sense of urgency to find effective intervention in the presymptomatic stage of AD so to reduce the risk of AD, delay or even prevent its onset. To more adequately diagnose different stages of the disease and especially in the early stage and predict future cognitive decline, computer-aided diagnostic classification is increasingly needed using biomarkers based on neuroimaging and other measurements. There is evidence that the pathogenic cascade of AD is thought to begin at least 1-2 decades prior to cognitive impairment, starting with accumulation of the amyloid-β1-42(Aβ1-42) plaques (Langbaum et al., 2013). Research has suggested that these early processes can be assessed using brain imaging and fluid biomarkers. Prior research work on Fluorodeoxyglucose (FDG) positron emission tomography (PET), Pittsburgh compound B (PIB), structural magnetic resonance imaging (sMRI) and functional measures of resting-state networks (rs-fMRI) has supported their validity as potential metabolic biomarkers. Among various neuroimaging techniques, ^{18}F-FDG PET characterizes the cerebral glucose hypometabolism related to AD and those at risk of AD. It even offers a reliable metabolic biomarker at pre-symptomatic stages. Despite major advances in FDG-PET used to track symptomatic patients, there is still a lack of sensitive, reliable, and accessible imaging algorithms capable of characterizing abnormal degrees of age-related metabolism decline in preclinical individuals at high risk for AD whom early intervention is most needed. Fig. 1 shows the different types of PET scans.

Recently minor cognitive impairment (MCI) in ^{18}F-FDG PET has been classified by a brain regional sensitivity mapping method based on summated index (Total Z score) by utilizing the sensitivity-distribution maps (Kakimoto et al., 2011).In other contemporary works a region of interest (ROI) mask is used to extract features and use incomplete random forest-robust support vector machine to perform classification (Lu et al., 2017). In general

for a classification algorithm based on 3D FDG-PET images the feature dimension is usually much larger than the number of subjects. Data with extremely high dimensionality has presented serious challenges to existing learning methods (Liu and Motoda, 2007; Friedman et al., 2001). With the presence of a large number of features, a learning model tends to overfit, affecting its performance.

In this context, when we apply three-dimensional statistical maps to do classification the feature dimension is usually much larger than the number of subjects, i.e., the so-called high dimension, small sample size problem. When a vast number of variables are measured from a small number of subjects, it is often necessary to reduce their high dimension features to low dimension features. In most cases, the information gets lost by mapping into a lower-dimensional space. However, the discarding information is always compensated by a more accurate space (or feature). There are two general approaches to performing dimensionality reduction: feature selection and feature extraction. Feature selection reduces the feature dimension by selecting a subset of original variables (Jain and Zongker, 1997). It aims to choose a small subset of the relevant features from the original ones according to certain relevance evaluation criterion, which usually leads to better learning performance for example higher learning accuracy for classification, lower computational cost, and better model interoperability (Tang et al., 2014). Feature extraction reduces the dimension based on mathematical projections, which transform the original features into a lower dimensional but more appropriate feature space (Guyon et al., 2008). There are some widely used algorithms in machine learning, e.g., principle component analysis (PCA) (Jolliffe, 2002), linear discriminant analysis (LDA) (Mika et al., 1999), as analytic tools for feature extraction.

The problem of a very large set of feature with respect to sample is called "the curse of dimensionality" by many. To address the problem of curse of dimensionality, we propose a patch based method of data collection for PET scans and on top of that we make a layer abstracting from a probablistic machine learning model for classification for a large set of ^{18}F-FDG PET images, and compare its performance. The approach is a empirical machine learning based model which includes patch based feature selection then maxpooling (Boureau et al., 2010) for feature agglomeration. We then form data vectors and apply probabilistic principle component analysis for feature extraction using dimension reduction

and finally AdaBoost (Rojas, 2009) for classification. The two systems serves as a comparison between the approaches for analyzing ^{18}F-FDG PET . Our goal is to discover an appropriate ^{18}F-FDG PET diagnosis system design by investigating their performance. We tested our hypothesis on the ADNI2 dataset across 668 subjects. We then carried out 10 fold cross validation in six different classification experiments comparing the two methods across various imaging measures.

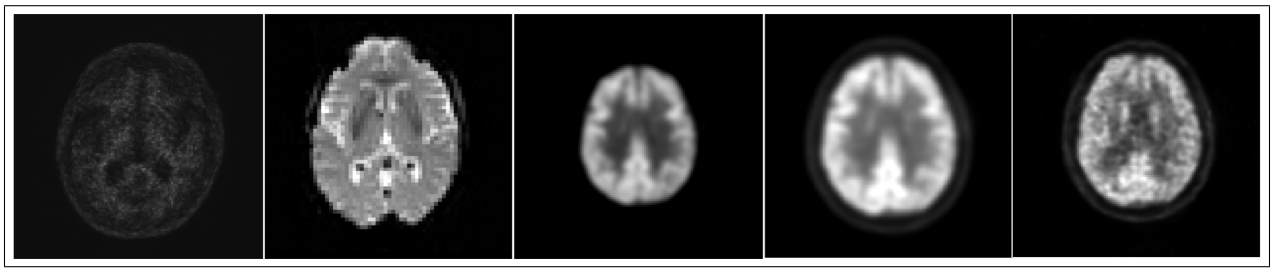

Figure 1: Different types of P.E.T Scans

2. Methods and Material

2.1. Theoretical Background

In prior work Zhang et al. (2016b) the authors study the surface deformation of the hippocampal region of the brain and use it as features for training a dictionary of basis vectors and associated sparse codes. This encoding learn sets of over complete basis vectors which are better able to capture structures and patterns inherent in the input data. The hippocampal surface is modeled from the hippocampal region of the brain and patches on its surface captures topological information effectively. Similar a system can be designed which uses a patch based feature extraction process and then use statistics derives form the patches for classification. However an important question for diagnostic classification based on voxel-based or surface-based maps is which statistics are best to analyze.

(Kakimoto et al., 2011) used the total z-score from the Brodmann Area sensitivity map in the brain surface, (Lu et al., 2017) calculated the mean voxel values from 116 VOIs, standard deviations of voxel values from the 116 anatomical VOIs, and mean voxel value differences between 54 pairs of the anatomical VOIs on left and right brain hemispheres. All VOIs were extracted form a AAL template similar to the one in Fig. **??** with more regions of varying

intensities. We hypothesize that a patch based extraction methods which selects overlapping volumes from each image in a number of samples and learns their reduced features can then be used for classification Algorithm. 1.

In this section, we briefly introduce the most relevant theoretical background in Dimensionality Reduction, Max-Pooling and Adaptive Boosting (AdaBoost)

Dimension Reduction Feature extraction approaches project features into a new feature space with lower dimensionality and the new constructed features are usually combinations of original features. Examples of feature extraction techniques include Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Singular Value Decomposition(SVD). Feature extraction maps the original feature space to a new feature space with lower dimensions by combining the original feature space. In that context principle component analysis (PCA) (Jolliffe, 2002) is a unsupervised machine learning algorithm widely used for dimensionality reduction. It used orthogonal transformations to convert a set of observations of possibly co-related values into a set of linearly uncorrelated variables called principle components.

After selecting the patches and structuring our 3D data into "sample $\times$ features", we would wish to analyze it summarizing its main characteristics. PCA is one of the most popular techniques for processing, compressing and visualising data. We use probabilistic PCA a variation of traditional PCA to reduce the dimensions of our selected features. Traditionally PCA's effectiveness is limited by its global linearity, to overcome that a combination of local linear PCA projections has been found to be able to capture data complexity efficiently. This model variant of PCA corresponds to the probability density unlike traditional PCA and enables it to combine PCA models (Tipping and Bishop, 1999). After reducing the dimensions of our dataset we will classify using AdaBoost.

Max Pooling State-of-the-art patch-based image representations involve a pooling operation that aggregates statistics computed from local descriptors. After obtaining features using convolution, we would next like to use them for classification. In theory, one could use all the extracted features with a classifier such as a softmax classifier, but this can be computationally challenging. We use max-pooling which summarizes the coded features over larger neighborhoods. To address this, first recall that we decided to obtain convolved fea-

tures because images have the "stationarity" property, which implies that features that are useful in one region are also likely to be useful for other regions. Thus, to describe a large image, one natural approach is to aggregate statistics of these features at various locations. If one chooses the pooling regions to be contiguous areas in the image and only pools features generated from the same (replicated) hidden units. Then, these pooling units will then be translation invariant. This means that the same (pooled) feature will be active even when the image undergoes (small) translations. Translation-invariant features are often desirable; in many tasks (e.g., object detection, audio recognition), the label of the example (image) is the same even when the image is translated. For example, if you were to take an MNIST digit and translate it left or right, you would want your classifier to still accurately classify it as the same digit regardless of its final position. Standard pooling operations include sum- and max-pooling. Sum-pooling lacks discriminability because the resulting representation is strongly influenced by frequent yet often uninformative descriptors, but only weakly influenced by rare yet potentially highly-informative ones. Max-pooling equalizes the influence of frequent and rare descriptors but is only applicable to representations that rely on count statistics, such as the bag-of-visual-words (BOV) and its soft-and sparse-coding extensions. Fig. 2 shows the whole process of max pooling. On the left hand side, pooling layer down-samples the volume spatially, independently in each depth slice of the input volume. On the right hand side, it is the most common max pooling shown with a stride of 2.

AdaBoost A number of statistical classifiers have been proposed for brain biomarker research. Support vector machine and Adaptive Boosting are the most popular ones. Adaboost short for "Abstract Boosting" is an approach to machine learning based on the idea of creating a highly accurate prediction rule by combining many relative weak and inaccurate rules. The AdaBoost Algorithm (Freund et al., 1996) was the first practical boosting algorithm, and remains one of the most widely used and studied, with applications in numerous fields. Adaboost can achieve more accuracy than any individual member classifier with unstable classifier. It can be used in conjunction with many other types of learning algorithms to improve their performance. One of the main ideas of the algorithm is to maintain a distribution or set of weights over the training set. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the weak learner is

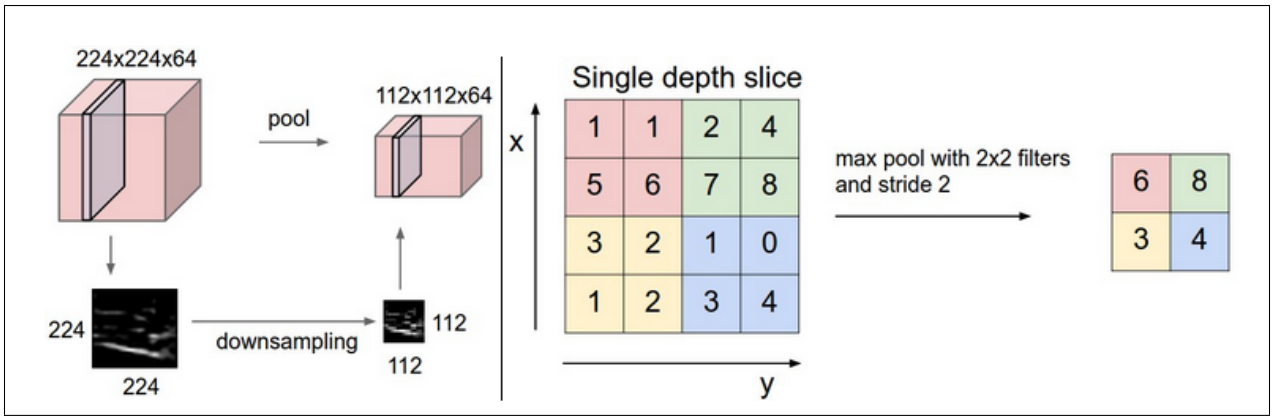

Figure 2: Pooling layer down-samples the volume spatially, independently in each depth slice of the input volume. Left: In this example, the input volume of size [25625616] is pooled with filter size 2; stride 2 into output volume of size [12812816]. Notice that the volume depth is preserved. Right: The most common down-sampling operation is max, giving rise to max-pooling, and here shown with a stride of 2, each max value is taken over 4 numbers (little 22 square).

forced to focus on the hard examples in the training set. Fig. 3 illustrates the general idea of the AdaBoost algorithm. The algorithm takes as input a training set $(x_1, y_1), \ldots, (x_m, y_m)$ where each x_i belongs to some domain or instance space X, and each label y_i is in some label set Y. For most of the discussion, we assume $Y = \{-1, +1\}$. Adaboost calls a given weak or base learning algorithm repeatedly in a series of rounds $t = 1, \ldots, T$. One of the main ideas of the algorithm is to maintain a distribution or set of weights over the training set. The weight of this distribution on training example i on round t is denoted as $D_t(i)$. Initially, all weights are set equally , but on each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training sets. The weak learner's job is to find a weak hypothesis $h_t : X \to \{-1, +1\}$ appropriate for the distribution D_t. The goodness of a weak hypothesis is measured by its error.

$$\epsilon_t = Pr_{i \sim D_t}[h_t(x_i) \neq y_i] = \sum_{i:h_t(x_i) \neq y_i} D_t(i)$$

Notice that the error is measured with respect to the distribution D_t on which the weak learner was trained. In practice, the weak learner may be an algorithm that can use the weights D_t on the training examples. Alternatively , when this is not possible, a subset of the training examples can be sampled according to D_t, and these (unweighted) resampled

examples can be used to train the weak learner Schapire (2013)

2.2. Computational Algorithms

This section introduces the computational algorithms of the proposed classification method in detail. Specifically, we build an integrated and automated framework to extract and analyze information from three dimensional ^{18}F-FDG PET scans, major algorithm is summarized in Alg. 1 and Fig. 5

2.2.1. FDG-PET acquisition and preprocessing using SPM (Statistical Parametric Mapping)

Screening and baseline FDG-PET scans were acquired for ADNI participants included in the current study (ADNI2) project via the Image Data Archive. Scans are performed within two weeks before or two weeks after the the in-clinical assessment at Baseline. A 30-min dynamic emission scan, consisting of four 5-min frames was acquired 30 to 60 minutes post-injection. The base frame image and the five co-registered frames (or all co-registered frames for the quantitative studies) are recombined into a co-registered dynamic image set. These image sets have the same image size (for example, 128×128×63) and voxel dimensions (for example, 2.0×2.0×2.0 mm) and remain in the same spatial orientation as the original PET image data. This is called native space. These files are uploaded to LONI in DICOM format. The scans are reoriented into a standard 160×160×96 voxel image grid having 1.5mm cubic voxels. with sections parallel to a horizontal section through the anterior and posterior commissures, normalize the images for individual variation in absolute image intensity, and apply a filter function. The images were uploaded to the Laboratory of Neuroimaging (LONI) ADNI website at UCLA and downloaded in NIFTI format.

Given a PET image, the alignment and the image segmentation are automatically performed using Statistical Parametric Mapping (SPM12) (Penny et al., 2011)[1]. First, all scans are aligned to a common space. Then, we borrow a brain mask from SPM, an AAL template to decide which regions to keep and which to remove. We align the template to the same space as we did to the test images and turn the AAL template into a mask by turning all the non-zero voxels to the value of "1". Third we compute the dot product of the FDG-PET

[1]http://www.fil.ion.ucl.ac.uk/spm/

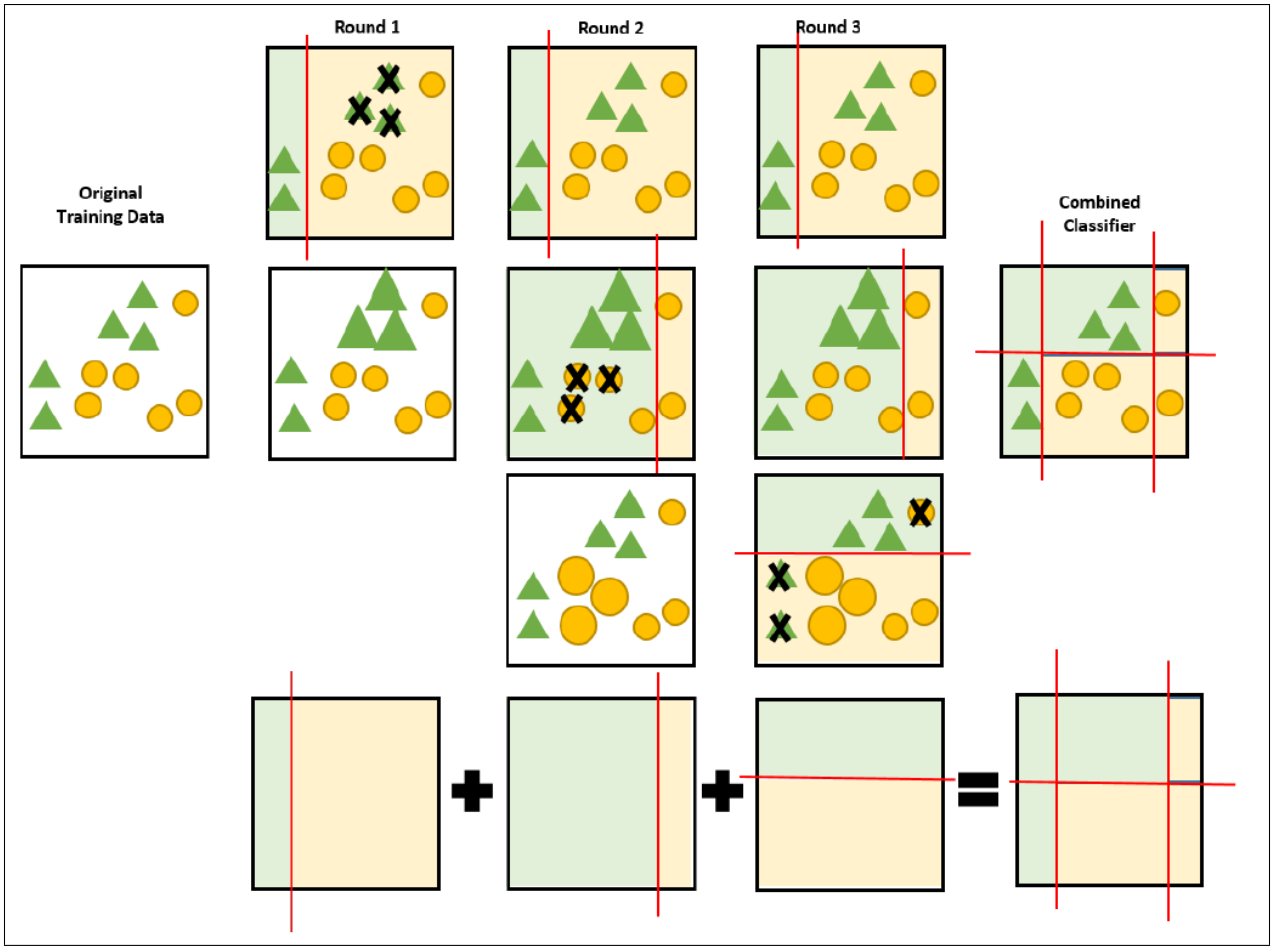

Figure 3: Illustration of the general idea of AdaBoost algorithm. The original training data is trained in three rounds. The first round is to create the first classifier, then the wrong classes will be given more weight and train in next round. In round 2, AdaBoost constructs a new classifier on the different weighted data. Similarly, it will give wrong classes higher weight and increase the probability of training in next round. Here, $\times$ represents wrong classes. In the third round, AdaBoost learns a new classifier based on last round weighted data. In this figure, the bigger shape means more weight to be trained. Finally, AdaBoost combines all classifiers in three rounds (calculated in last line) into the final classifier.

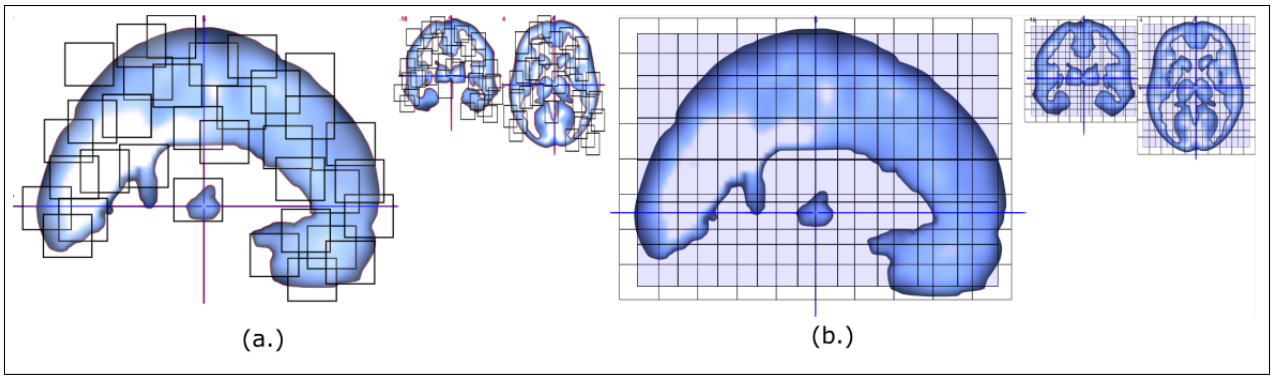

Figure 4: (a). Unstructured patches are generated randomly and ensure overlapping by creating enough of them. (b). Structured patches with overlap shown in a slightly stronger shade of blue.

image scans and the mask generated to segment the region of interest. Fourth, we conduct spatial smoothing with a Gaussian kernel of the full width at half maximum (FWHM) equal to (8; 8; 8) in three directions (x; y; z).

2.2.2. Patch Generation and Image Representation

After we identify the region of interest as the cerebral cortex in the FDG-PET scans we are left with a $80 \times 95 \times 80$ voxel intensities which represent the metabolic activity. For this study we choose the entire cerebral cortex as the basis for comparison or as we would call it the biomarker. The feature dimension with the FDG-PET data is much larger than the number of subjects and is therefore prone to overfitting. We first randomly generate a number of small $10 \times 10 \times 10$ windows on each image volume to obtain a collection of small image patches with different amounts of overlap. The procedure is in fact equivalent to applying a high-pass filter to the original volume. As a result, the region of interest (ROI) are still present, but some low frequency signals have disappeared.

With respect to three dimensional data one can argue as to which patch orientation and overlapping will give the best results and will not compromize the perfomance of the training algorithm. As such the processing pipeline thus has a large number of tunable parameters, such as the patch size or the orientation or even the choice of step size in the training process. It turns out that getting these parameters set correctly can make a major difference in the performance of practical applications. In fact, these parameteres have a greater contribution towards the classification performance than the training algorithm. We

make two sets of patches 1.) Structured overlapping patches with uniform overlap and 2.) Unstructured overlapping patches with random overlap.

Illustration of the patch generation procedure is shown in Fig.4 (a,b).

2.2.3. Feature Selection and Extraction

After we identify the region of interest as the cerebral cortex in the FDG-PET scans we are left with a $80 \times 95 \times 80$ voxel intensities which represent the metabolic activity. For this study we choose the entire cerebral cortex as the basis for comparison or as we would call it the biomarker. The feature dimension with the FDG-PET data is much larger than the number of subjects and is therefore prone to overfitting. We first randomly generate a number of small $10 \times 10 \times 10$ windows on each image volume to obtain a collection of small image patches with different amounts of overlap. The procedure is in fact equivalent to applying a high-pass filter to the original volume. As a result, the region of interest (ROI) are still present, but some low frequency signals have disappeared.

In the pipeline we pool and arranged the patches into a linear array $X_i = (x_i^1, \ldots, x_i^n)$ where n is the number of feature and $i \in \{1 \ldots m\}$, where m is the number of samples in a group and stack them for the concerned group $X = (X_1, \ldots, X_i, \ldots, X_m)$. The corresponding labels are also generated in this step $Y = (y_1, \ldots, y_m)$. We will then try to find out the linear and non-linear separability of the data using machine learning algorithms.

Illustration of the patch selection procedure is shown in Fig.5 (c).

At this step we have a good understanding of our input data for the training process, theoretically we could the data matrix X for classification. However when a vast number of variables are measured from a relatively small dataset, the feature dimension is usually much larger than the sample size and the volume of the space increases so fast that the available data becomes sparse. In such a problem an enormous amount of data is required to ensure that there are several samples with each combination of values thus ensure no overfitting. With a fixed number of training samples, the predictive power reduces as the dimensionality increases. To overcome this issue we choose to effectively represent our data by reducing the dimensions of our feature space. In machine learning, dimensionality reduction is the introduction of new feature space where the original features are represented. The new space

is of lower dimension that the original space. e.g., principle component analysis (PCA) (Jolliffe, 2002), linear discriminant analysis (LDA) (Mika et al., 1999).

A major limitation of traditional PCA is that it is non-parametric as there is no probabilistic model for observed data. In 1999 ME Tipping (Tipping and Bishop, 1999) proposed a probabilistic PCA model (PPCA) in which the principle axes of a set of observed data vectors may be determined through maximum-likelihood estimation. The main idea of PPCA is based on the latent variable models which seeks to relate a d-dimensional observed data vector $\mathbf{t}$ to a corresponding q-dimensional vector of latent variable (i.e. the unobserved variables that can be inferred from the model) x.

$$\mathbf{y} = Wx + \mu + \epsilon \tag{1}$$

where, latent variables $x \sim \mathcal{N}(0, I)$). the noise model is described as $\epsilon \sim \mathcal{N}(0, \Psi)$, and the $(d \times q)$ parameters matrix W contains the factor loading (Factor Analysis). μ is the mean, given this formulation the observational vectors $\mathbf{t}$ are also normally distributed $\mathbf{t} \sim \mathcal{N}(\mu, C)$. In regards to equation (4.1) the probability model for the case of isotropic noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is

- $y|x \sim \mathcal{N}(WX + \mu, \sigma^2 I)$
- $y \sim \mathcal{N}(\mu, C_y)$, where $c_y = WW^T + \sigma^2 I$ (where C_y is the covariance matrix for the observed data y)

The Maximum Likelihood solution for PPCA is obtained as:
$W_{MLE} = U_q(\Lambda_q - \sigma^2_{MLE} I)^{1/2} R$,
where U_q is a matrix of q leading principal directions(eigen values of the covariance matrix), Λ_q is a diagonal matrix of corresponding eigenvalues.
$\sigma^2_{MLE} = \frac{1}{d-q} \sum_{j=q-1}^{d} \lambda_j$ represents the variance lost in the projection and R is an arbitrary $q \times q$ rotation matrix (corresponding to rotations in the latent space).

A key motivation for this model is that because of the diagonality of Ψ (the variance not accounted by the latent factors ergo noise.) observed variables are conditionally independent given latent factors x. The intention is that the dependencies between the data variables t

are explained by a smaller number of latent variables x, while ϵ represents variance unique to each observation variable, unlike PCA which treats covariance and variance similarly.

Illustration of the PPCA is shown in Fig.5 (d).

The choice of classifier to process the lower dimension data is AdaBoost, as we saw in Section 2.1, it is a good classifier for biomarker research along with Support vector machine and a few others, we will compare Adaboost with other classifiers and prove that it is stable and robust classifier for biomarker research in the early detection of Alzheimer's .

Algorithm 1: Patch Based Feature Extraction & Dimensionality Reduction Pipeline

Input: A number of FDG-PET images

Output: A set of features and labels for the six classification experiment

The segmented dataset is divided and segregated into 6 binary parcels.

for *exp* ← { *[AD vs. CU], [AD vs. EMCI], [AD vs. LMCI], [CU vs. EMCI], [CU vs. LMCI], [EMCI vs. LMCI].* } **do**

- n ← number of samples
- p ← number of patches
- Initialize X = [n × p]
- **for** *sample* ∈ *exp* **do**
  - A three dimensional patch of 10 × 10 × 10 is created in order to extract meaning full information from the 3D segmented PET scan.
  - A number of patches are generated over the PET volume and overlapping is ensured
  - All patch windows are max-pooled and linearly arranged into a vector B(Fig.5 (b)).
  - X_{sample} ← B. Where sample ∈ m
- **end**
- In X the high dimensionality of the data is reduced by probabilistic PCA to avoid over fitting (Fig. 5(c)).
- classification of X: sample × features, Y: labels, is performed by AdaBoost.

end

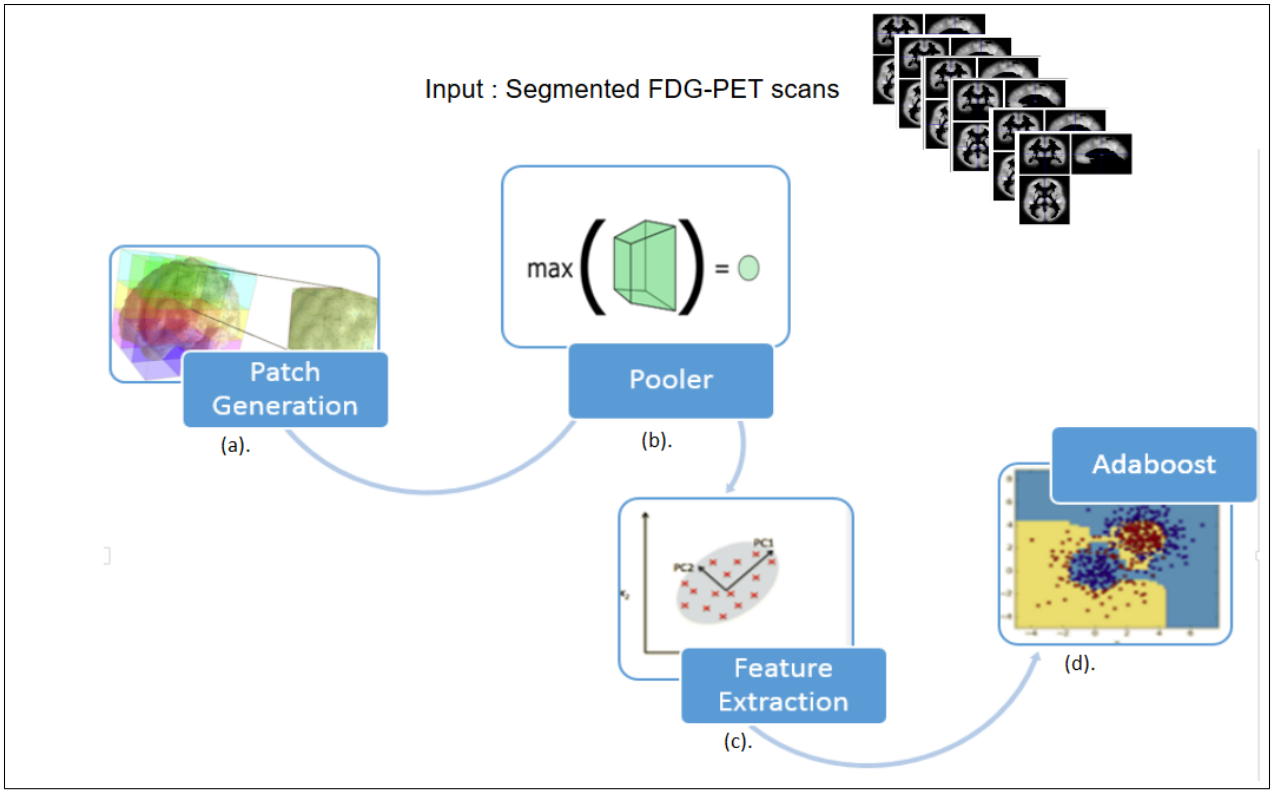

Figure 5: Pipelines for patch based feature extraction, (a). patches are generated for feature extraction. (b). patches are pooled to obtain specific activation. (c). the pooled values are linearly arranged and PCA is used to reduce it to a lower dimension space. (d). Adaboost is applied on the new feature space.

2.3. Subjects

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

In this study, we studied a total of 668 subjects in the ADNI2 baseline dataset, within this population, there were 146 who had Alzheimer's (AD), 158 had impairment (LMCI), 178 had early mild cognitive impairment (EMCI) and 186 were normal control (CN).

Table 1: Demographic Information of 668 Subjects in the ADNI2 Baseline Dataset.

	Male	Female	Age	Min / Max Age	APOE1	APOE2	FAQ
AD	85	61	74.73 ± 8.15	56 / 90	3.11	3.63	13.39
LMCI	84	74	72.5 ± 7.5	55 / 91	3.03	3.54	03.62
EMCI	102	76	71.3 ± 7.2	55 / 89	2.94	3.42	02.08
CN	89	97	73.5 ± 6.25	57 / 89	2.86	3.24	00.16

3. Experimental Results

We design six experiments (1). AD vs. CU (2). AD vs. EMCI (3). AD vs. LMCI (4). CU vs. EMCI (5). CU vs. LMCI (6). EMCI vs. LMCI. The objective here is to independently study the inter-cohort relationships in hope of learning more about the class separation in AD clinical group.

We run the six experiments over the pipeline described in Algorithm 1. An N-fold cross validation protocol was adopted to estimate classification accuracy. All subjects were randomly divided into N folds. The surface biomarkers were selected by training on N-1 folds and the test was performed on the remaining fold. We rotated this procedure for N times to estimate the accuracy. In this paper, we choose N = 10 to complete the classification.The output of each classification experiment was compared to the ground truth, and a contingency table was computed to indicate how many class labels were correctly identified as members of one of the two classes. The rows of the contingency table represent the true classes and the columns represent the assigned classes. The cell at row r and column c is the number of subjects whose true class is r while their assigned class is c. Natural performance measures for classification problems mainly based on error rate or accuracy. However, higher accuracy does not necessarily imply better performance on target task. In two-category classification, one method for handling c-class problem is to consider c 2-class problems: $\frac{\omega_i}{not\ \omega_i}$ (Fawcett, 2004). Therefore, confusion matrix was proposed as a method to measure classifier performance. TP, FP, TN, FN represents number of true positives, number of false positives, number of true negatives and number of false negative, respectively. The matrix in Table. 2 represents a possible combination of ground truth and predicted classification for two classes.

	Assigned Class	
True Class	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Table 2: The Confusion Matrix

The total of TP, FP, FN, TN refers to the total number in the classification. Five performance measures F_1 Score, Recall, Specificity, Positive predictive value and Negative predictive value were calculated as follows:

$$F_1 \text{ Score} = 2 \times \frac{\text{TP}}{\text{FN} + 2 \times \text{TP} + \text{FP}}.$$

In a given population precision measures the amount of true cases classified correctly and recall the strength of that number. When you have both high precision and high recall it would mean that the population is well classified. F_1 Score is the measure that is the harmonic mean of precision and recall. The reason for taking harmonic mean is because it is more appropriate when dealing rates and ratios.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Recall in this context is also referred as the true positive rate or sensitivity. is a statistical measure of how well a binary classification test correctly identifies a condition and the probability of correctly labeling members of the target class.

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}.$$

The specificity is a statistical measure of how well a binary classification test correctly identifies the negative cases.

$$\text{Positive predictive value} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

Where positive predictive value (PPV) is also referred to as precision, which measures the probability of a positive prediction is correct.

$$\text{Negative predictive value} = \frac{\text{TN}}{\text{TN} + \text{FN}}.$$

which measures the probability of a negative prediction is correct.

All these measures provide relevant information about the classification and no single measure tells the entire story. For example consider a scenario where 90% of the population does not have a disease and the 10% population is misclassified by the classifier, the accuracy would still be 90%. Thus we should use multiple measures.

There are some standard performance evaluation measures for classification study. Bigger values usually mean stronger classification power. We also computed the area-under-the-curve (AUC) of the receiver operating characteristic (ROC). The ROC is the average value of sensitivity for all the possible values of specificity. Such an index is especially useful in a comparative study of two diagnostic tests. If two tests are to be compared, it is desirable to compare the entire ROC curve rather than at a particular point (Swets, 1979). The maximum AUC=1 means that the diagnostic test is perfect in the differentiation between the diseased and stable. This happens when the distribution of test results for the diseased and stable do not overlap. AUC = 0.5 means the chance of discrimination that curve located on diagonal line in ROC space.

3.1. Variety of Features

In the training matrix we then involve additional features which enables us to make well defined classification, we include the genetic information and Functional Activities Questionnaire (FAQ) Scores even Age and Gender were involved. the two alleles of Apolipoprotein E(APOE) are available, APOE genotype is represented by combination of $e2$ (episilon 2),$e3$ and $e4$. Each individual will have one of the following combinations: $e2/e2$, $e2/e34$, $e2/e44$, $e3/e3$, $e3/e4$, $e4/e4$. Although the order between two genotypes for each person doesn't matter, they are represented in the order of $e2 \geq e3 \geq e4$. After feature annotation, the dimension of the dataset was reduced to a reasonable size and classification was performed.

We ran the pipeline in Sec. 2.2.3 over 1.) Voxels information 2.) Voxel information and genetic scores 3.) Voxel, genetic scores and other demographic information. The results of the experiment are shown Table. 3

Exp.	Voxels	Voxels + Gene + FAQ	Voxels + Gene + FAQ + Demo
AD vs. CU	88.97	95.10	96.30
AD vs. EMCI	76.92	81.69	83.68
AD vs. LMCI	63.26	69.44	73.42
CU vs. EMCI	54.49	62.63	69.94
CU vs. LMCI	63.87	82.19	78.28
EMCI vs. LMCI	58.04	59.47	57.38

Table 3: Comparison results between different sets of features. The measure used in F1-Score, three different sets of features are used to compare the effectiveness of Voxel, Apoe$E1$, Apoe$E2$, FAQ and age/gender in classification.

It is clear from the Table. 3 that the combination of voxel information and other genetic and demographic features gives the best classification result. The groups with high class seperation AD vs. CU has the maximum F-1 score of $\sim 96\%$.

3.2. Comparing Fearure learning algorithms

We wonder how other dimension reduciton algorithms react to our method, we compare probablistic PCA with frequently used learning methods such as linear Singular Valued Decomposition and non linear Kernel-PCA. As indicated in Table. 4 the F_1 Score of AD vs CU is best when the dimensions are reduced using singular valued decomposition (SVD) with 97% F_1 Score. The high Recall and Specificity implies AD vs CU is well classified. Again we believe the ^{18}F-FDG PET biomarker along with Apoe$E2$ and Apoe$E3$ the two alleles of Apolipoprotein E and the Functional Activities Questionnaire (FAQ) are extremely sensitivity if classifying between Alzheimer's and Normal Control. Column two in Table.4 shows that the subjects in AD vs EMCI are separable to a good extent with F_1 Score of 83.3% in case of Principle Component Analysis (PCA) and a Recall of 84.5% and Specificity of 85.7% shows that both the classes have been evenly separated. Kernel PCA (Mika et al., 1998) performed the best in remaining experiments. With CU vs LMCI the group separation is of one i.e. the disease progression there is one stage in between CU and LMCI and we

Measure	Method	AD CU	AD EMCI	AD LMCI	CU EMCI	CU LMCI	LMCI EMCI
PCA	F1 Score	96.85	83.33	73.42	69.94	78.28	57.38
	Recall	94.38	84.50	75.00	67.50	78.07	58.04
	Specificity	99.26	85.71	75.00	58.90	74.52	52.46
	PPV	99.46	82.19	71.91	72.58	78.49	56.74
	NPV	92.46	87.64	77.85	63.48	74.40	53.79
SVD	F1 Score	97.09	81.85	70.23	66.67	77.62	57.47
	Recall	95.33	85.18	68.62	67.78	77.83	58.82
	Specificity	98.56	83.59	72.19	65.59	73.58	53.01
	PPV	98.92	78.76	71.91	65.59	77.41	56.17
	NPV	93.83	88.76	69.62	67.41	74.05	55.69
Kernel PCA	F1 Score	96.65	82.47	76.65	71.46	80.01	60.39
	Recall	95.77	82.75	78.01	68.47	78.06	59.56
	Specificity	96.50	85.47	77.91	70.80	77.02	54.90
	PPV	97.31	82.19	75.34	74.73	82.25	61.23
	NPV	94.52	85.95	80.37	64.04	72.78	53.16

Table 4: Classification Results with PCA, SVD and Kernel PCA. In this comparison we used Adaboost as a fixed classifier for all the reduction technique.

observe in column 5 of Table. 4 the F_1 Score is 80%. With AD vs LMCI and CU s EMCI with Kernel PCA the F_1 Score is 76.65% and 71.46% respectively. The last experiment (LMCI vs EMCI) F_1 Score of 60.3% is the most difficult to classify many good and popular classifiers failed to classify the complex nature of early Mild impairment (EMCI) and impairment (LMCI). EMCI and LMCI are derived from the Mild Cognitive Impairment (MCI) stage in ADNI1 and the participants reported a subjective memory concern the difference in EMCI and LMCI is decided by the Wechsler Memory Scale Logical (WMS). We believe there is no clear separation with ^{18}F-FDG PET as our biomarker and maybe more specific ROI may lead us to a more concrete conclusion.

3.3. Comparing Classifiers

To demonstrate the effectiveness of AdaBoost as a good classifier for biomarker related research in Alzheimer's classification we compare nine classifiers and report the best four as illustrated in Table. 5. The F_1 Score of AD vs CU is best for Gaussian Process (GP) with a 97.3% F_1 Score compared to 96.2% in AdaBoost. 98.38% PPV and 95.20% NPV indicates both the classes have been effectively classified. The class separation is maximum in case of AD vs CU in reference to disease progression that explains the high F_1measure. For AD vs EMCI, AD vs LMCI, CU vs EMCI and CU vs LMCI the F_1Score is highest case of Gaussian Process. In case of CU vs LMCI AdaBoost performs comparatively poor in comparison to other classifiers. For EMCI vs LMCI the NPV is poor for GP and Linear SVM which means one class is poorly classifier and the classification accuracy is inconsistent. In this case AdaBoost performed good as it gave a more consistent NPV and PPV the recall and sensitivity is also the best in this experiment. For further comparison between dimensionality reduction we keep our classifier as AdaBoost because of a more correct classification when AdaBoost was used as classifier across all the experiments.

Measure	Method	AD CU	AD EMCI	AD LMCI	CU EMCI	CU LMCI	LMCI EMCI
Nearest Neighbor	F1 Score	96.53	83.33	75.98	73.65	84.87	60.77
	Recall	95.76	88.46	79.69	67.41	77.67	56.52
	Specificity	96.50	84.02	76.60	75.00	90.00	52.71
	PPV	97.31	78.76	72.60	81.18	93.54	65.73
	NPV	94.52	91.57	82.91	58.98	68.35	43.03
Linear SVM	F1 Score	96.79	83.27	77.35	75.79	84.95	67.27
	Recall	96.27	86.67	78.72	65.87	77.43	56.75
	Specificity	96.52	84.65	78.52	82.14	90.67	59.74
	PPV	97.31	80.13	76.02	89.24	94.08	82.58
	NPV	95.20	89.88	81.01	51.68	67.72	29.11
Gaussian Process	F1 Score	97.34	83.68	77.58	75.96	85.64	63.76
	Recall	96.31	86.76	80.74	68.69	79.35	55.93
	Specificity	97.88	85.10	78.10	79.10	89.68	54.00
	PPV	98.38	80.82	74.65	84.94	93.01	74.15
	NPV	95.20	89.88	83.54	59.55	71.51	34.17
AdaBoost	F1 Score	96.85	83.33	73.42	69.94	78.28	57.38
	Recall	94.38	84.50	75.00	67.50	78.07	58.04
	Specificity	99.26	85.71	75.00	58.90	74.52	52.46
	PPV	99.46	82.19	71.91	72.58	78.49	56.74
	NPV	92.46	87.64	77.85	63.48	74.40	53.79

Table 5: Classification results between different classifiers. Popular classifiers are used to perform analysis on the best feature set.

3.4. Comparing Dimensionality Reduction with Sparse Coding

This experiment further investigates what if some other feature learning method is used insted of dimensionality reduction. Sparse coding has been applied in many fields like audio

processing, text mining and image recognition, Sparse coding concerns the problem of reconstruting data vectors using sparse combination of basis vectors. (Lin et al., 2014; Olshausen and Field, 1996). We report the variation in the results in Figure. 6.

Measure	Method	AD CU	AD EMCI	AD LMCI	CU EMCI	CU LMCI	LMCI EMCI
F1 Score	FE	96.53	83.33	75.98	73.65	84.87	60.77
	SCC	96.53	83.33	75.98	73.65	84.87	60.77
Recall	FE	95.76	88.46	79.69	67.41	77.67	56.52
	SCC	95.76	88.46	79.69	67.41	77.67	56.52
Specificity	FE	96.50	84.02	76.60	75.00	90.00	52.71
	SCC	95.76	88.46	79.69	67.41	77.67	56.52
PPV	FE	97.31	78.76	72.60	81.18	93.54	65.73
	SCC	95.76	88.46	79.69	67.41	77.67	56.52
NPV	FE	94.52	91.57	82.91	58.98	68.35	43.03
	SCC	95.76	88.46	79.69	67.41	77.67	56.52

Table 6: Classification results between different classifiers. Popular classifiers are used to perform analysis on the best feature set.

Area under the Curve is used to measure receiver operating characteristic (ROC) curve. The area is a measure of the predictive power of the classification experiment to classify a randomly chosen positive example more accurately than a randomly chosen negative example. ROC is the graphical plot between the true positive rate and the false positive rate, it was first used during World War 2 for detecting enemy objects from friendly ships and noise. Know as Signal detection theory, it measures the ability of radar receiver operators to identify enemy ships.

Fig. 6 illustrates the classification in all clinical groups with different features performance comparison with receiver operating characteristic (ROC) curves and area under curve (AUC) measures. The results for Patch based Feature Extraction (PFE) and Patch based Feature Extraction (PFE) with other demographic features are computed with the proposed pipeline

in Algorithm 1, patch based Dictionary Learning (PDL) is applied to dataset before training as described in Appendix .1 and Section 4.1.1. Within the four statistics, the result of using Feature Extraction with Voxles, APOE and FAQ are better than others. Among all AUC measures, PFE+Demo feature achieved the best performance (AUC = 0.99).

In this comparison we plot the AUC and compare both of our methods with and without demographics. The graphs show the superior accuracy of including rich features over only the regional voxel intensities.

4. Discussion

In this paper, we propose a new imaging prediction system called patch based feature extraction for Alzheimer's disease diagnosis via ^{18}F-FDG PET analysis (PFE) for the early prediction of cognitive impairment and diagnose AD and its various stages in progression (EMCI and LMCI). Deriving motivation from some prior work in surface morphometry statistics (Zhang et al., 2016b,a) we selected a series of different extent overlapped patches and applied max-pooling. For dimension reduction, we selected a series of algorithms and performed meta analysis to determine the best performing one. We base our classifier as AdaBoost as it has proven to be a excellent choice for biomedical research (**?**). The result on 663 baseline subjects from ADNI show that our PFE framework performed better than some other standard image measures. Our study has two main findings. First, the combined statistics for each subject, consisting of APOE gene information, FAQ scores and other demographic features carries rich information on distributed metabolic activity, which measures volumetric deformation and are also applicable for prediction and classification research. The newly combined statistics practically encodes a great deal of information that would be otherwise overlooked, volumetric based computer-aided diagnostic research is more powerful by analyzing these rich features. Second, we try to demonstrate the feasibility of learning algorithms in image based prediction and classification research. SCC is an efficient sparse coding algorithm and its superior computational efficiency helps us to account for biological differences and achieve deep learning. (Gregor and LeCun, 2010) pointed out that sparse coding is a special case of sparse encoder method (Vincent et al., 2010; Baldi, 2012). These deep architectures have shown to lead to state-of-the-art results on a number of challenging classification and regression problem. We find out the the encoding of features involving volumetric data is large thus costly to implement and there is no significant performance improvement in the training.

Both Feature extraction and feature selection are capable of improving learning performance, lowering computational complexity, building better generalizable models, and decreasing required storage. Feature extraction maps the original feature space to a new feature space with lower dimensions by combining the original feature space. It is difficult to link the features from original feature space to new features. Therefore further analysis of new fea-

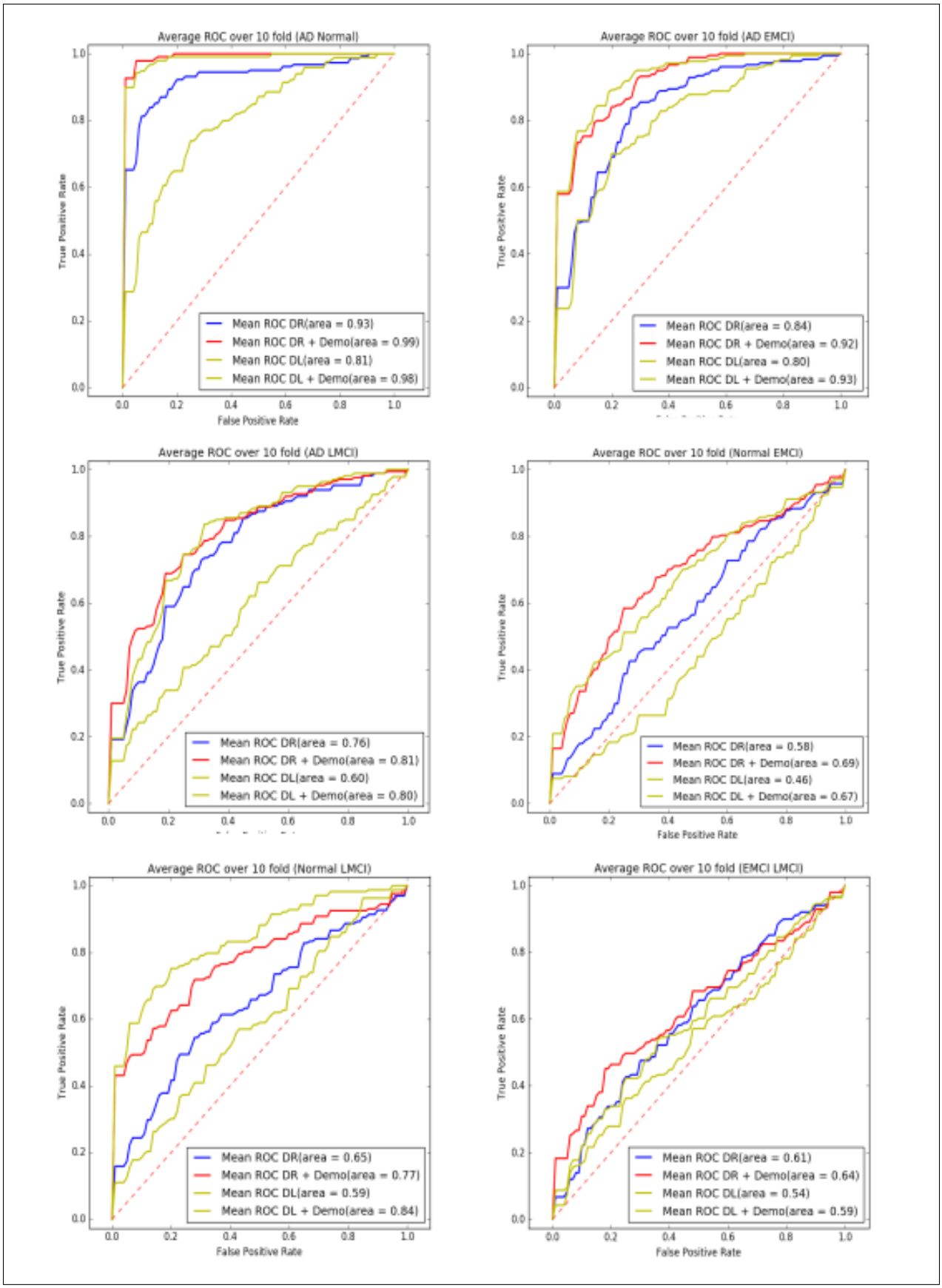

Figure 6: ROC comparing all clinical groups

tures is problematic since there is no physical meaning for the transformed features obtained from feature extraction techniques. While feature selection selects a subset of features from the original feature set without any transformation, and maintains the physical meanings of the original features. In this sense, feature selection is superior in terms of better readability and interpretability. This property has its significance in many practical applications such as finding relevant genes to a specific disease and building a sentiment lexicon for sentiment analysis. Typically feature selection and feature extraction are presented separately. Via sparse learning such as $\ell 1$ regularization, feature extraction (transformation) methods can be converted into feature selection methods (Masaeli et al., 2010). In order to analyze images more efficiently, a dictionary that allows us to represent them as a superposition of a small number of its elements so that we can reduce each image to a small number of its coefficients (Schnass and Vandergheynst, 2008). Similarly sparse coding (Lin et al., 2014) has been proposed to use a small number of basis vectors (also called dictionary) to represent local features effectively and concisely and help image content recognition. From the input image data, sparse coding learns an over-complete set of basis vectors (dictionary), which have been used to select the most germane features (Friedman, Hastie and Tibshirani, 2001). To learn local imaging features, image patches are usually selected to form dictionaries. Dictionary learning and sparse coding (Mairal et al., 2009) has shown to be efficient for many tasks such as image deblurring (Yin et al., 2008), image super-resolution functional connectivity analysis (Lv et al., 2015a,b), and image classification. In many computer vision, medical imaging and bioinformatics applications (Mairal et al., 2009; Moody et al., 2012; Lv et al., 2015b) dictionary learning and sparse coding leads to state-of-the-art results.

4.1. Methods for ADNI classification using PET subsectioncans

The development of automatic methods for the accurate classification of patients into clinical groups from imaging data has been the aim of a number of ADNI studies. Thus PET classification plays an important role in medical image retrieval, which is a part of decision making in medical image analysis. ADNI 2 PET data is new compared to ADNI 1 and has two different modalities for traditional MCI i.e. EMCI and LMCI which are the early and late stages in impairment.

For 3D PET scans "Z-score" images are used to represent the active parts of the brain it will show all pixels with values below the lower threshold in blue, and pixels above the upper threshold in red (Ishii, 2014). PET scans have also been analyzed by utilizing adjusted T statistics and an automated voxel-based procedure (Herholz et al., 2002) and Machine Learning algorithms to address the high dimensionality of the statistical maps (Illán et al., 2011; Higdon et al., 2004). Recently minor cognitive impairment (MCI) in ^{18}F-FDG PET has been classified by a brain regional sensitivity mapping method based on summated index (Total Z score) by utilizing the sensitivity-distribution maps (Kakimoto et al., 2011). In other contemporary works a region of interest (ROI) mask is used to extract features and use incomplete random forest-robust support vector machine to perform classification (Lu et al., 2017). In previous work within our lab MRI images were classified using surface measures of ventricular enlargement and sparse coding then applied on the 2D-patch features (Zhang, Shi, Stonnington, Li, Gutman, Chen, Reiman, Caselli, Thompson, Ye et al., 2016a; Zhang, Stonnington, Li, Shi, Bauer, Gutman, Chen, Reiman, Thompson, Ye et al., 2016b) with 96.7% accuracy. These images were functional MRIs and the features were a combination of surface statistics, we build our idea on a similar model we fist design a empirical machine learning based model. Using three dimensional patches (i.e., small sub volumes of the image defined as three-dimensional cubes) we extract information. A very similar 3D patch based feature selection is described in (Coupé et al., 2011), In this work, voxels with similar surrounding neighborhoods are considered to belong to the same structure and thus are used to estimate the final label. Our data is also along the same lines.

4.1.1. Finding Dictionary and Sparse Codes

Dictionary learning is widely used in machine learning, neuroscience, signal processing, and statistics. It is the learning of the basis set, also called the dictionary, to adapt it to specific data, an approach that has recently proven to be very effective for signal processing in the audio and image processing domains. Different from traditional feature extraction methods like principle component analysis and its variants, sparse coding learns non-orthogonal and over-complete dictionaries which have more flexibility to represent data. Stochastic Coordinate Coding has been used successfully in the past (Lin et al., 2014; Mairal et al.,

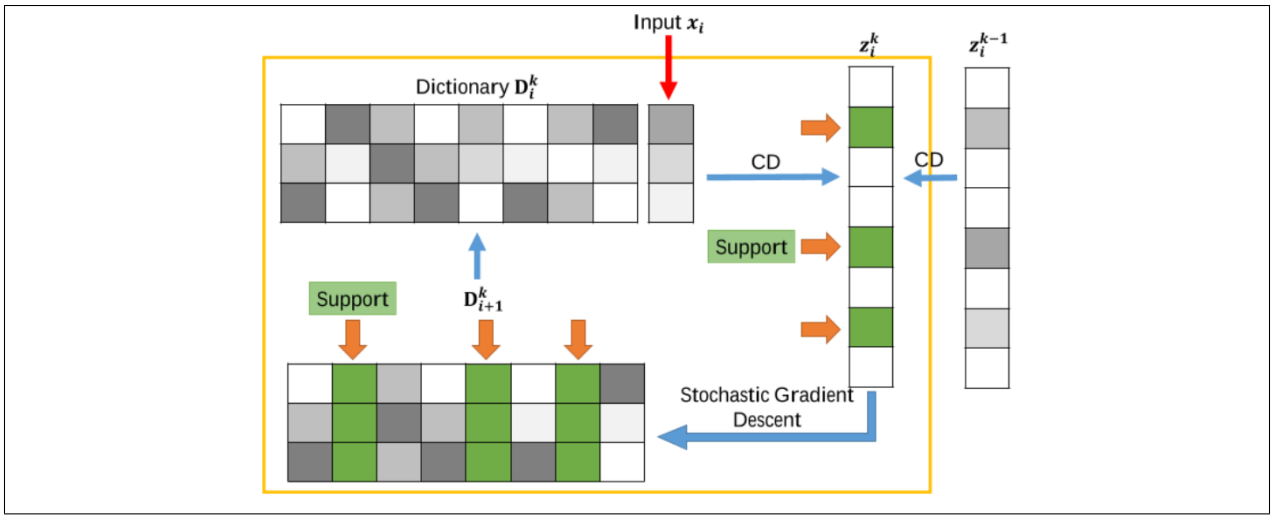

Figure 7: Figure shows one iteration of the sparse coding algorithm. (a).Take an image patch x_i. (b). Perform few steps of coordinate descent to find the support (non zero entries) of the sparse code. (c). Update the support of the dictionary by second order stochastic gradient descent to obtain a new dictionary.

2009).

Given a data set $\mathbf{X} = (x_1 \dots x_n)$ of image patches, each image patch is a p-dimensional vector i.e., $x_i \in \mathbb{R}^p$, $i = 1, \dots, n$. Specifically, suppose there are m atoms $\mathbf{d_j} \in \mathbb{R}, j = 1, \dots, m$, where the number of atoms is usually much smaller than the number of image patch p. Each patch can be represented as $\mathbf{x_i} = \sum_{j=1}^{m} z_{i,j} d_j$. In this way, the p-dimensional vector $\mathbf{x_i}$ is represented by an m-dimensional vector $\mathbf{z_i} = (z_{i,1}, \dots, z_{1,m})^T$ which means the learned feature vector $\mathbf{x_i}$ is a sparse vector.

As illustrated in Fig. 7 shows the input x_i and the sparse vector z_i^k where i is the sample number and k is the iteration.

5. Conclusion and Future Work

In this paper, we present a frameworks that combines three dimensional voxel statistics with machine learning to deal with high dimensional features before classification. We also used dictionary learning and sparse coding to study the effectiveness of leature learning algorithm with respect to the original framework. We applied AdaBoost to classify different AD stages. Our comprehensive experiments showed the effectiveness of patch based methods in three dimensional Positron Emission Tomography (PET). We obtained 96% classification

with voxel + demographic statistics. Our proposed PFE method performs well in experiments with high group separation. With experiments having low group separation both the algorithms struggled. This indicates that there is significant metabolic change in AD vs. CU but in other stages the metabolic change may be unpredictable to some extent.

We hope our work sheds light on the utilization of PET images as biomarker information in classifying Non-AD classes with a greater accuracy. It also invokes the use of varied multiple features in the diagnosis of Alzheimer's via some clinical group classification. In the future, we plan to apply our systems to other cortical and sub-cortical regions in the brain, more specifically to design a better ROI based feature selection method so as to identify regions in the cortex and sub-cortex responsible for cognitive decline.

Acknowledgements

This work was partially supported by the National Institutes of Health (R21AG043760 to JS, WZ, RJC and YW, R21AG049216 to WZ and YW, RF1AG051710 and U54EB020403 to YW, R01AG031581 and P30AG19610 to RJC) and the National Science Foundation (DMS-1413417 to YW, IIS-1421165 to WZ and YW).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to Rev December 5, 2013 support ADNI clinical sites in Canada. Private sector contributions

are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix .1. ***Stochastic Coordinate Coding***

The following section describes Stochastic Coordinate Coding (SCC) algorithm (Lin et al., 2014). The dictionary is initialized via any initialization method and denoted as D_1^1. The sparse codes are initialized as $z_i^0 = 0$ for $i = 1, \ldots, n$. Where i is the index of data points and the superscript denotes the number of epoch. The algorithm is describes as follows:

(1). Get an image patch x_i.

(2). Update z_i^k via one or a few steps of coordinate descent:

$$z_i^k = CD(D_i^k, z_i^{k-1}, x_i). \tag{.1}$$

Specifically, for j from 1 to m, update the j-th coordinate $z_{i,j}^{k-1}$ of z_i^{k-1} cyclically as follows:

$$b_j \leftarrow (d_{i,j}^k)^T(x_i - D_i^k z_i^{k-1}) + z_{i,j}^{k-1}, z_{i,j}^{k-1} \leftarrow h_\lambda(b_j), \tag{.2}$$

where h is the soft thresholding shrinkage function. We call such cycle as one step of coordinate descent. The updated sparse code is then denoted by z_i^k.

(3). Update the dictionary D by using stochastic gradient descent:

$$D_{i+1}^k = P_{B_m}(D_i^k - \eta_i^k \Delta_{D_i^k} f_i(D_i^k, z_i^k)) \tag{.3}$$

where P denotes the projection operator. We set the learning rate as an approximate of the inverse of the Hessian matrix. The gradient of D_i^k can be obtained as follows:

$$\Delta_{D_i^k} f_i(D_i^k, z_i^k) = (D_i^k z_i^k - x_i)(z_i^k)^T. \tag{.4}$$

(4). $i = i + 1$. If $i \geq n$, then set $D_1^{k+1} = D_{n+1}^k$, $k = k + 1$ and $i = 1$.

When the data sets are very large, the learning rate η_i^k will be very small. In this case, the dictionary will not change very much and the efficiency of the training will decrease. In

practice tuning the learning rate is tricky and sensitive. To obtain the learning rate we use the Hessian matrix of the objective function. It is shown that the following matrix provides an approximation of the Hessian: $\mathbf{H} = \sum_{k,i} z_i^k (z_i^k)^T$, when k and i go to infinity. According to the second order stochastic gradient descent, we should use the inverse matrix of the Hessian as the learning rate. However, computing a matrix inversion problem is computationally expensive. In order to get the learning rate, we simply use the diagonal elements of the matrix H. The matrix H is updated as follows:

$$H \leftarrow H + z_i^k (z_i^k)^T. \tag{.5}$$

Alg. 2 summarizes the steps described in section 4.1.1

Algorithm 2: SCC (Stochastic Coordinate Coding)

Input: Data set $X = (x_1, x_2, \ldots, x_n) \in R^{p \times n}$, ensure $D \in R^{p \times m}$ and $Z = (z_1 \ldots z_n) \in R^{m \times n}$

Output: $D = D_n^k$ and $z_i = z_i^k$ for $i = i, \ldots, n$.

Initialize $D_1^1, H = 0$ and z_i^0 for $i = 1, \ldots, n$,
for $k = 1$ *to* k *do* **do**
  for $i = 1$ *to* n **do**
    Get an image patch x_i
    Update z_i^k via one or a few steps of coordinate descent:
    $z_i^k \leftarrow CD(D_i^k, z_i^{k-1}, x_i)$.
    Update the Hessian matrix and the learninig rate:
    $H \leftarrow H + z_i^k (z_i^k)^T,\ \eta_{i,j}^k = \frac{1}{h_{jj}}$.
    Update the support of the dictionary via SGD:
    $d_{i+1,j}^k \leftarrow d_{i,j}^k - \eta_{i,j}^k z_{i,j} (D_i^k z_i^k - x_i)$
    If $i = n$, set $D_1^{k+1} = D_{n+1}^k$.
  end
end

References

Baldi, P., 2012. Autoencoders, unsupervised learning, and deep architectures, in: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp. 37–49.

Boureau, Y.L., Ponce, J., LeCun, Y., 2010. A theoretical analysis of feature pooling in visual recognition, in: Proceedings of the 27th international conference on machine learning (ICML-10), pp. 111–118.

Burns, D.A., Iliffe, S., . Enfermedad de Alzheimer , 338, b158.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54, 940–954.

Fawcett, T., 2004. Roc graphs: Notes and practical considerations for researchers. Machine learning 31, 1–38.

Freund, Y., Schapire, R.E., et al., 1996. Experiments with a new boosting algorithm, in: icml, pp. 148–156.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. volume 1. Springer series in statistics Springer, Berlin.

Gregor, K., LeCun, Y., 2010. Learning fast approximations of sparse coding, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 399–406.

Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A., 2008. Feature extraction: foundations and applications. volume 207. Springer.

Herholz, K., Salmon, E., Perani, D., Baron, J., Holthoff, V., Frölich, L., Schönknecht, P., Ito, K., Mielke, R., Kalbe, E., et al., 2002. Discrimination between alzheimer dementia and controls by automated analysis of multicenter fdg pet. Neuroimage 17, 302–316.

Higdon, R., Foster, N.L., Koeppe, R.A., DeCarli, C.S., Jagust, W.J., Clark, C.M., Barbas, N.R., Arnold, S.E., Turner, R.S., Heidebrink, J.L., et al., 2004. A comparison of classification methods for differentiating fronto-temporal dementia from alzheimer's disease using fdg-pet imaging. Statistics in medicine 23, 315–326.

Illán, I., Górriz, J.M., Ramírez, J., Salas-Gonzalez, D., López, M., Segovia, F., Chaves, R., Gómez-Rio, M., Puntonet, C.G., Initiative, A.D.N., et al., 2011. 18 f-fdg pet imaging analysis for computer aided alzheimers diagnosis. Information Sciences 181, 903–916.

Ishii, K., 2014. Pet approaches for diagnosis of dementia. American Journal of Neuroradiology 35, 2030–2038.

Jain, A., Zongker, D., 1997. Feature selection: Evaluation, application, and small sample performance. IEEE transactions on pattern analysis and machine intelligence 19, 153–158.

Jolliffe, I., 2002. Principal component analysis. Wiley Online Library.

Kakimoto, A., Kamekawa, Y., Ito, S., Yoshikawa, E., Okada, H., Nishizawa, S., Minoshima, S., Ouchi, Y., 2011. New computer-aided diagnosis of dementia using positron emission tomography: brain regional sensitivity-mapping method. PloS one 6, e25033.

Langbaum, J.B., Fleisher, A.S., Chen, K., Ayutyanont, N., Lopera, F., Quiroz, Y.T., Caselli, R.J., Tariot, P.N., Reiman, E.M., 2013. Ushering in the study and treatment of preclinical alzheimer disease. Nature Reviews Neurology 9, 371–381.

Lin, B., Li, Q., Sun, Q., Lai, M.J., Davidson, I., Fan, W., Ye, J., 2014. Stochastic coordinate coding and its application for drosophila gene expression pattern annotation. arXiv preprint arXiv:1407.8147 .

Liu, H., Motoda, H., 2007. Computational methods of feature selection. CRC Press.

Lu, S., Xia, Y., Cai, W., Fulham, M., Feng, D.D., Initiative, A.D.N., et al., 2017. Early identification of mild cognitive impairment using incomplete random forest-robust support vector machine and fdg-pet imaging. Computerized Medical Imaging and Graphics .

Lv, J., Jiang, X., Li, X., Zhu, D., Zhang, S., Zhao, S., Chen, H., Zhang, T., Hu, X., Han, J., et al., 2015a. Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function. IEEE Transactions on Biomedical Engineering 62, 1120–1131.

Lv, J., Lin, B., Zhang, W., Jiang, X., Hu, X., Han, J., Guo, L., Ye, J., Liu, T., 2015b. Modeling task fmri data via supervised stochastic coordinate coding, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 239–246.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., 2009. Online dictionary learning for sparse coding, in: Proceedings of the 26th annual international conference on machine learning, ACM. pp. 689–696.

Masaeli, M., Dy, J.G., Fung, G.M., 2010. From transformation-based dimensionality reduction to feature selection, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 751–758.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.R., 1999. Fisher discriminant analysis with kernels, in: Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop., IEEE. pp. 41–48.

Mika, S., Schölkopf, B., Smola, A.J., Müller, K.R., Scholz, M., Rätsch, G., 1998. Kernel pca and de-noising in feature spaces., in: NIPS, pp. 536–542.

Moody, D.I., Brumby, S.P., Rowland, J.C., Gangodagamage, C., 2012. Unsupervised land cover classification in multispectral imagery with sparse representations on learned dictionaries, in: Applied Imagery Pattern Recognition Workshop (AIPR), 2012 IEEE, IEEE. pp. 1–10.

Olshausen, B.A., Field, D.J., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607.

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., 2011. Statistical parametric mapping: the analysis of functional brain images. Academic press.

Rojas, R., 2009. Adaboost and the super bowl of classifiers a tutorial introduction to adaptive boosting. Freie University, Berlin, Tech. Rep .

Schapire, R.E., 2013. Explaining adaboost, in: Empirical inference. Springer, pp. 37–52.

Schnass, K., Vandergheynst, P., 2008. Dictionary learning based dimensionality reduction for classification, in: Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on, IEEE. pp. 780–785.

Swets, J.A., 1979. Roc analysis applied to the evaluation of medical imaging techniques. Investigative radiology 14, 109–121.

Tang, J., Alelyani, S., Liu, H., 2014. Feature selection for classification: A review. Data Classification: Algorithms and Applications , 37.

Tipping, M.E., Bishop, C.M., 1999. Mixtures of probabilistic principal component analyzers. Neural computation 11, 443–482.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research 11, 3371–3408.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., et al., 2013. The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. Alzheimer's & Dementia 9, e111–e194.

Yin, W., Osher, S., Goldfarb, D., Darbon, J., 2008. Bregman iterative algorithms for \ell_1-minimization with applications to compressed sensing. SIAM Journal on Imaging sciences 1, 143–168.

Zhang, J., Shi, J., Stonnington, C., Li, Q., Gutman, B.A., Chen, K., Reiman, E.M., Caselli, R., Thompson, P.M., Ye, J., et al., 2016a. Hyperbolic space sparse coding with its application on prediction of alzheimers disease in mild cognitive impairment, in: International

Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 326–334.

Zhang, J., Stonnington, C., Li, Q., Shi, J., Bauer, R.J., Gutman, B.A., Chen, K., Reiman, E.M., Thompson, P.M., Ye, J., et al., 2016b. Applying sparse coding to surface multivariate tensor-based morphometry to predict future cognitive decline, in: Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on, IEEE. pp. 646–650.