# Selecting and Extracting Effective Features for Automated Diagnosis of Alzheimer's Disease

Mohamed M. Dessouky
Faculty of Electronic Engineering, Computer Science and Engineering Dep.
Menoufiya University

Mohamed A. Elrashidy, Taha E. Taha
Faculty of Electronic Engineering, Computer Science and Engineering Dep.
Menoufiya University

Hatem M. Abdelkader
Faculty of Computers and Information
Menoufiya University

## ABSTRACT

In this paper, a Computer Aided Diagnosis (CAD) system is proposed to provide a comprehensive analytic method for extracting the most significant features of Alzheimer's disease (AD). It consists of three stages: feature selection, feature extraction, and classification. This proposal selects the features that have different intensity level at all images and discarding the features that have the same intensity level to reach the fewer subset of features that have the most impact distinctive of AD. Then reduces the features by proposing a new feature extraction algorithm that minimizes intra separately distance of AD features. Finally, a Linear Support Vector Machine (SVM) classifier was used to perform binary classifications among AD patients. The data set that used for testing the proposed model consists of 120 cross-sectional Structural MRI images from the Open Access Series of Imaging Studies (OASIS) database. Experiments have been conducted on Open Access Series of Imaging Studies (OASIS) database. The results show that the highest classification performance is obtained using the proposed model, and this is very promising compared to Principle Component Analysis (PCA) and Linear Discriminate Analysis (LDA).

## General Terms

Pattern Recognition, Diagnosis, Alzheimer's disease, and Computer Aided Diagnosis

## Keywords

Feature Extraction, Feature Selection, Support Vector Machine, Principle Component Analysis, and Linear Discriminate Analysis.

## 1. INTRODUCTION

Dementia is a syndrome due to disease of the brain, usually chronic, characterized by a progressive, global deterioration in intellect including memory, learning, orientation, language, comprehension and judgment.

Alzheimer's disease (AD) is a one of the most important example of dementia which mostly affects people over 65 years old and whose incidence rate grows exponentially with age, almost doubling in every 5 years. Still, apart from a few exceptions, the factors that trigger the onset of AD remain unknown. It is a progressive disease, this means that it worsens over time, and for which there is currently no cure, leading eventually to death. The very early stages are often mistakenly confused with the normal process of ageing or linked to stress and it is often characterized by episodic losses of short term memory and difficulty to grasp new ideas. Structural and Functional neuroimaging allow studying of brain pathology at macro and micro molecular level like Magnetic Resonance Imaging (MRI), single photon emission computed tomography (SPECT), and Positron emission tomography (PET), table 1 shows different types of neuro-imaging . It summarizes different types of Neuroimaging techniques, where CT = computerized tomography; MRI = magnetic resonance imaging; fMRI = functional magnetic resonance imaging; DTI = diffusion tensor imaging; SPECT = single-photon emission computed tomography; PET = positron emission tomography; MEG = magneto-encephalography; BOLD = blood oxygen level-dependent; NAA = N-acetyl aspartate; Cr = Creatinine; AD = Alzheimer's disease.

The early detection of Alzheimer's disease still a challenge because of the estimation of the scans depends on manual directing and visual reading. This preclinical stage is also known as Mild Cognitive Impairment (MCI). As the brain damage progresses, other cognitive deterioration appear and the disease becomes obvious. In the late stages, persons are completely dependent on caregivers even for the most basic daily tasks such as eating, bathing or dressing [1-4]. Till now, there is no cure for Alzheimer's disease, but its early detection is important to a successful treatment, slacken the progression of symptoms. So, the development of automatic diagnostic tools, which use as major sthece of information 3D images of the brain, has attracted great attention in last years. Computer Aided Diagnosis (CAD) allows early detection of the disease at early stages, and structural brain images are useful in this task.

Computer Aided Diagnosis (CAD) tools have been successfully applied in the AD detection using the analysis of particular features in a functional brain image. The Fisher Discriminant Ratio (FDR) was used to choose only the most discriminant voxels. Then, the resultant features were projected onto a low dimensional subspace using a decomposition technique called NMF. At last, a modified SVM with bounds of confidence was utilized as the classifier [5]. In [6] a CAD system was designed that consists of three stages: voxel selection, feature extraction and classification. Voxels are chose in terms of their significance, by using Mann–Whitney–Wilcoxon U-Test. Then, Factor Analysis is used to do the feature reduction step, by separating mutual factors and factor loadings from the chosen voxels. Finally, a Linear Support Vector Machine (SVM) classifier is used to execute clustering of the input images. In [7] a built CAD system that consist of the group of voxels specifying the antecedents and consequences of the Association Rules (ARs) are chosen as input voxels for posterior dimensionality reduction. Feature extraction is defined by a next reduction of the chosen voxels using principal component analysis (PCA) or partial least squares (PLS) techniques while classification is done by a support vector machine (SVM).

**Table 1. Typical findings of different brain imaging methods used in AD and MCI diagnosis.**

| Neuroimaging Technique | Finding |
|---|---|
| CT | Tissue Atrophy |
| MRI | Tissue Atrophy. It is more specific in grey matter |
| fMRI | Changes in blood oxygenation level (BOLD signal) |
| DTI | Connectivity and organization in white matter. |
| Spectroscopy | Chemical content of the brain, such as NAA/Cr ratio. |
| SPECT | Changes in cerebral perfusion |
| PET | Changes in glucose metabolism |
| MEG | Measure magnetic fields and get information about brain electrical activity. |

The most two important feature extraction algorithms that has been used by different previous works are Principal Component Analysis (PCA) [20, 21], and Linear Discriminant Analysis (LDA) [22]. The main advantage of the PCA and LDA is that they are able to combine the input features during the process of dimensionality reduction, while ranking methods only look at one feature at a time. The main disadvantage of these techniques is the higher computational needs. So, we tried to overcome these weakness in the proposed technique.

In this paper, Voxel-Based Morphometry is used by the package of VBM8 [10] with the SPM8 package [11] for analyzing, preprocessing and normalizing the images. Then feature selection step was made by selecting the voxels that have the intensity level different at all image and neglecting the voxels that have the same intensity level at all images. Next, a proposed feature extraction approach is tested to extract an effective voxels from images and compared with Principle Component Analysis (PCA) and Linear Discriminate Analysis (LDA). Finally, Linear Support Vector Machine (SVM) classifier is used for clustering. Section 2 talks about Feature extraction and selection, PCA and LDA, section 3 talks about the proposed Algorithm, then section 4 presents the experimental results and lastly the conclusion, and result discussion.

## 2. DIMENSION REDUCTION

### There are two types of dimensionality reduction:

**1. Feature Selection**: Selecting a subset of the existing features without a transformation.

**2. Feature Extraction:** Transforming the existing features into a lower dimensional space. As shown in Figure 1.

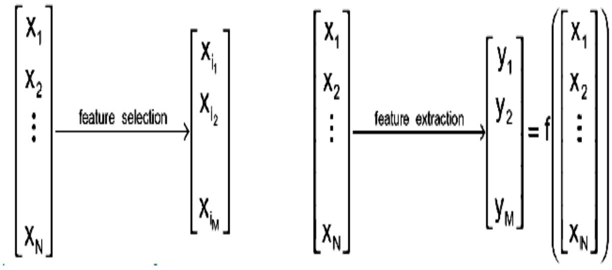
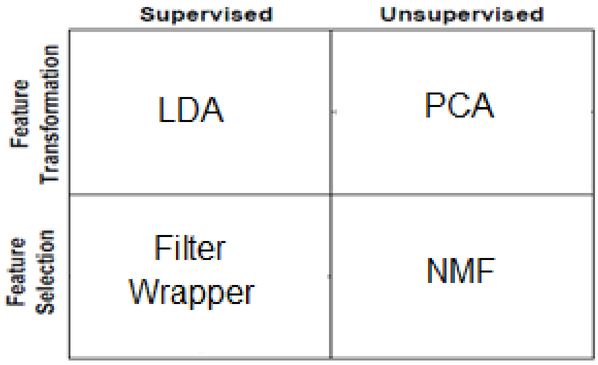

$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{feature selection}} \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_M} \end{bmatrix}$
$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{feature extraction}} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}\right)$

**Fig 1: Feature Selection and Extraction**

The two key distinctions in dimension reduction research are the distinction between supervised and unsupervised techniques and the distinction between feature transformation and feature extraction techniques. The dominant techniques are feature subset selection and principal component analysis. As shown in Figure 2.



| | Supervised | Unsupervised |
|---|---|---|
| Feature Transformation | LDA | PCA |
| Feature Selection | Filter<br>Wrapper | NMF |

**Fig 2: Feature Extraction and selection**

## 2.1 Feature Selection

Feature selection (FS) algorithms occupy another approach to dimension reduction by finding the “best” least subset of the original features, without transforming the data to a new set of dimensions. For the purpose of knowledge discovery, interpreting the output of algorithms based on feature extraction can often prove to be intricate, as the transformed features may have no physical meaning to the domain expert. In the other hand, the dimensions retained by a feature selection procedure can generally be directly explained. Feature selection in the context of supervised learning is a reasonably well posed problem. The objective can be to determine features that are correlated with or predictive of the class label. Or more comprehensively, the objective may be to select features that will construct the most accurate classifier. In unsupervised feature selection the object is less well posed and consequently it is a much less explored area [14].

In supervised learning, selection techniques typically incorporate a search strategy for exploring the space of feature subsets, including methods for deciding a suitable starting point and generating successive elected subsets, and an evaluation criterion to rate and compare the candidates, which serves to guide the search process. The evaluation schemes used in both supervised and unsupervised feature selection techniques can generally be divided into three broad categories:

### 2.1.1 Filter

Filter approaches attempt to eject irrelevant features from the feature set before the application of the learning algorithm. Initially, the data is analyzed to select those dimensions that are most relevant for describing its structure. The selected feature subset is then used to train the learning algorithm. Feedback regarding an algorithm's performance is not needed during the selection process, though it may be useful when trying to measure the effectiveness of the filter [14].

### 2.1.2 Wrapper

Wrapper methods for feature selection make use of the learning algorithm itself to select a set of pertinent features. The wrapper makes a search through the feature space, evaluating selected feature subsets by estimating the predictive accuracy of the classifier built on that subset. The aim of the search is to find the subset that maximizes this criterion [14].

### 2.1.3 Embedded

Embedded approaches apply the feature selection process as a complete part of the learning algorithm. The most example of this are the decision tree building algorithms. There are a number of neural network algorithms that also have this characteristic, e.g. Optimal Brain Damage [14].

## 2.2 Feature Extraction

Involves the production of a new set of features from the original features in the data, through the application of some mapping. Well-known unsupervised feature extraction methods include Principal Component Analysis (PCA) [14, 16] .The important corresponding supervised approach is Linear Discriminant Analysis (LDA) [14, 17]. The famous feature transformation technique is Principal Components Analysis (PCA) that transforms the data into a reduced space that captures most of the variance in the data. PCA is an unsupervised technique in that it does not take class labels into account. By contrast Linear Discriminant Analysis (LDA) seeks a transformation that maximizes between-class separation [14].

### 2.2.1 Principal Component Analysis (PCA)

PCA is known as the best data representation in the least-square sense for classical recognition. It is commonly used to decrease the dimensionality of images and get most of information. The central idea behind PCA is to find an orthonormal set of axes pointing at the direction of maximum covariance in the data. It is often used in representing facial images. The idea is to find the orthonormal basis vectors, or the eigenvectors, of the covariance matrix of a set of images, with each image treated as a single point in a high-dimensional space. It is supposed that the facial images form a connected sub region in the image space. The eigenvectors map the most significant variations between faces and are preferred over other correlation techniques that assume that every pixel in an image is of equal importance [16]. PCA is a powerful tool for analyzing data and once you have found these patterns in the data, and you compress the data by reducing the number of dimensions, without much loss of information.

**Methods:**

Step 1: Get some data.

Step 2: Subtract the mean.

Step 3: Calculate the covariance matrix.

Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix.

Step 5: Choose components and form a feature vector.

Step 6: Derive the new data set.

### 2.2.2 Linear Discriminate Analysis (LDA)

LDA is used to make the feature extraction and to classify samples of unknown classes based on training samples with known classes. It get a linear transformation of k-dimensional samples into an m-dimensional space (m < k), so that samples pertinence to the same class are close together, but samples from different classes are far apart from each other. This method maximizes the ratio of between-class variance to within-class variance in any data set; thereby, the theoretical maximum separation in the linear sense will be guaranteed. Since LDA require directions that are efficient for discrimination, it is the optimal classifier for specializing classes that are Gaussian distribution and have equal covariance matrices. LDA requires a transformation matrix that in some sense maximizes the ratio of the between-scatter matrix to the within-scatter matrix. The within-scatter matrix is defined as [17]

$$S_w = \sum_{j=1}^{K} \sum_{i=1}^{N_j} (y_i^j - \mu_j)(y_i^j - \mu_j)^T \qquad (1)$$

Where $y_i^j$ is the ith sample of class j, $\mu_j$ is the mean of class j, K is the number of classes, and $N_j$ is the number of samples in class j. The between-scatter matrix is defined as [17]

$$S_b = \frac{1}{K} \sum_{i=1}^{K} (\mu_i - \mu)(\mu_i - \mu)^T \qquad (2)$$

Where µ is the mean of all classes. The goal is to maximize the between-class measure and to minimize the within class measure [17]

$$\text{(i.e., max. } \frac{\det|S_b|}{\det|S_w|} \text{ )} \qquad (3)$$

To achieve this, we can find a transformation Matrix [17]

$$W_{opt} = \arg max_w \left(\frac{W^T S_b W}{W^T S_w W}\right) = [w_1 w_2 \dots w_m] \qquad (4)$$

Where $\{w_i\}$ is the set of generalized eigenvectors of $S_b$ and $S_w$ .Once the transformation is found, the classification problem is simply a matter of finding the class whose transformed mean is closest to the transformed testing image. PCA and LDA are two important feature extraction methods and have been widely applied in a variety of areas [17].

## 3. PROPOSED APPROACH

This paper presents a proposed automated CAD system that can automatically diagnosis the Alzheimer's disease. The OASIS data set [12] with different 120 subject aged 18 to 96 years is used. Some of them have been clinically diagnosed with very mild to moderate Alzheimer's disease and others were negative and have no Alzheimer disease. Most two important feature extraction methods (PCA and LDA) was tested with the proposed algorithm on the subjects. In order to compare all feature extraction algorithms, they were all tested with Voxel Intensity (VI) features. Classification was carried out by a Support Vector Machine (SVM) with linear kernel. Parameter optimization was performed within a nested Cross Validation (CV) procedure. Figure 3, gives the proposed approach used for feature extraction and selection algorithm. Figure 4, illustrates the proposed algorithm flow chart.

**Pseudo-code for the proposed Algorithm:**

```
1- Read the MRI images.
2- Make Preprocessing and Normalization for these images.
3- Get the brain shape features.
4- Select features from brain shape features using proposed feature selection method.
5- Extract special features from selected feature using PCA, LDA, and proposed extraction method.
6- Use cross validation for training and testing the results.
7- Do classification using Linear Support Vector Machine (SVM) classifier.
8- Show results.
```

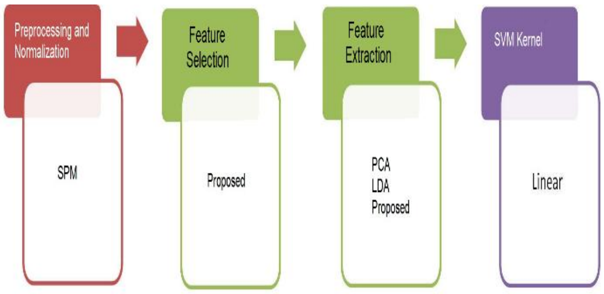


**Fig 3: Proposed approach to compare feature extraction algorithms**

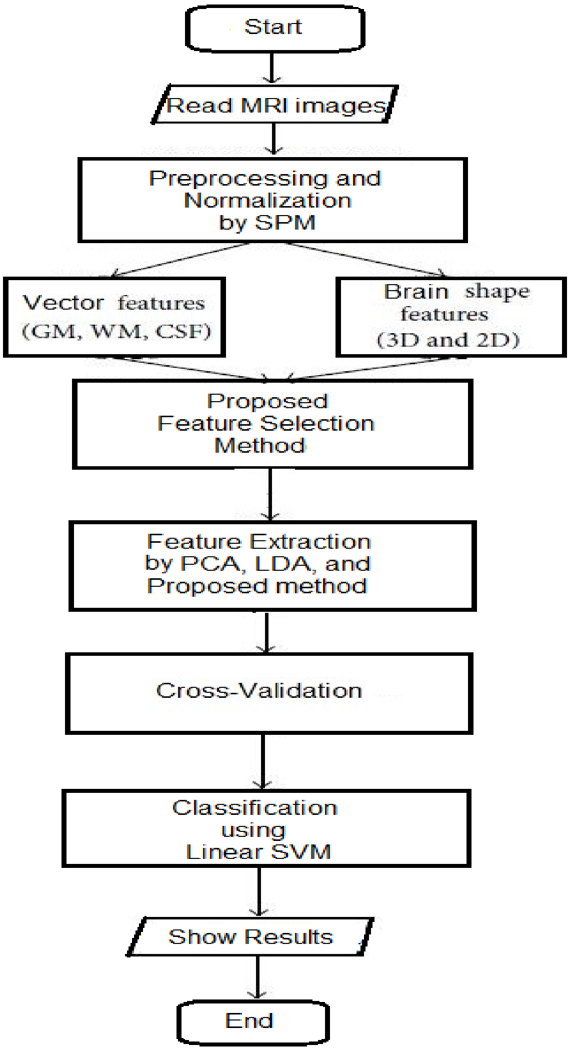

Start
Read MRI images
Preprocessing and Normalization by SPM
Vector features (GM, WM, CSF)
Brain shape features (3D and 2D)
Proposed Feature Selection Method
Feature Extraction by PCA, LDA, and Proposed method
Cross-Validation
Classification using Linear SVM
Show Results
End

**Fig 4: Proposed algorithm flow chart**

## 3.1 Preprocessing and Normalization of the database

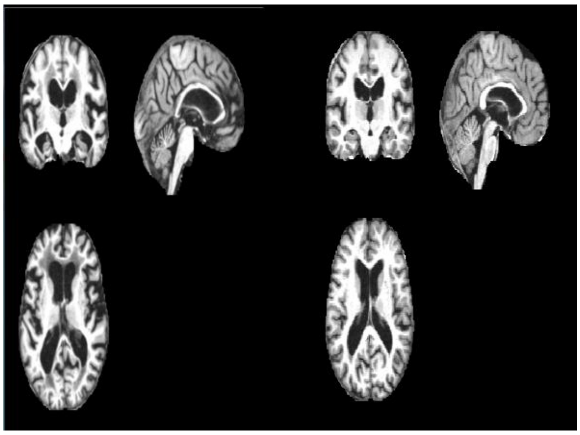
One hundred and twenty subjects of male and female (aged 18 to 96 yrs.) were selected from the Open Access Series of Imaging Studies (OASIS) database [12]. OASIS data set consists of a cross-sectional collection of 416 subjects covering the adult life span aged 18 to 96 including individuals with early-stage Alzheimer's Disease (AD). For the study we get 120 subjects there are 49 subjects who have been diagnosed with very mild to mild AD and 71 nondemented. A summary of subject demographics and dementia status is shown in table 2.

**Table 2.Summary of subject demographics and dementia status.**

| | Very mild to mild AD | Normal |
|---|---|---|
| No of subjects | 49 | 71 |
| Age | 63-96 | 33-94 |
| Education | 1-5 | 1-5 |
| socioeconomic status (SES) | 1-5 | 1-5 |
| Clinical Dementia Rating (CDR) | 0.5 – 1 – 2 | 0 |
| Mini-Mental State Examination (MMSE) | 16-30 | 25-30 |

Note: Education codes correspond to the following levels of education: 1 less than high school grad., 2: high school grad., 3: some college, 4: college grad., 5: beyond college. Categories of socioeconomic status: from 1 (biggest status) to 5 (lowest status). MMSE score ranges from 0 (worst) to 30 (best).

In this study, Neuroimaging data of Open Access Series of Imaging Studies (OASIS) was used; a series of magnetic resonance imaging data sets that is publicly available for study and analysis. One hundred and twenty cross-sectional subjects of male and female (aged 18 to 96 yrs.) were selected. For each subject, three or fthe individual T1-weighted magnetic resonance imaging scans obtained in single imaging sessions are included. Multiple within-session acquisitions provide extremely high contrast-to-noise ratio, making the data amenable to a wide range of analytic approaches including automated computational analysis. Multiple (three or fthe) high-resolution structural T1-weighted magnetization prepared rapid gradient echo (MP-RAGE) images, there were acquired on a 1.5-T Vision scanner in a single imaging session. Image parameters: TR= 9.7 msec, TE= 4.0 msec, Flip angle= 10, TI= 20 msec, TD= 200 msec, 128 sagittal 1.25 mm slices without gaps. All photos are 3-D and its dimensions are 176 X 208 X 176 voxels size. As shown in Figure 5.



**(a) Demented and mild with AD subject**

**(b) Nondemented with AD subject**

**Fig 5: Demented and nondemented subjects with AD**

The analysis of structural magnetic resonance images is done by Voxel based morphometric (VBM) approaches which allows between- and within-groups comparison of grey and white matter volume or density [8,9]. VBM is well suited for large-scale cross-sectional and longitudinal studies that examine normal age-related neuro-morphologic change. In a typical neuro-morphologic study of aging, structural magnetic resonance images (MRI) are acquired, spatially normalized to common stereotactic coordinates, and segmented into grey matter, white matter, and cerebrospinal fluid (CSF). Statistical parametric maps (SPM) are generated that reflect differences between or within groups (or the relationship with a continuous variable) in each individual voxel. [9]

Voxel based morphometric (VBM) approaches to the analysis of structural magnetic resonance images allow for between- and within-groups comparison of grey and white matter volume or density. Voxel Based Morphometry (VBM) involves a voxel-wise comparison of the local concentration of gray matter between two groups of subjects. The procedure is relatively straightforward and involves spatially normalizing high-resolution images from all the subjects in the study into the same stereotactic space. This is followed by segmenting the gray matter from the spatially normalized images and smoothing the gray-matter segments. Voxel-wise parametric statistical tests which compare the smoothed gray-matter images from the two groups are performed. Corrections for multiple comparisons are made using the theory of Gaussian random fields. [8,9]

This paper describes the steps involved in VBM, with particular emphasis on segmenting gray matter from MR images with non-uniformity artifact. The evaluations of the assumptions that underpin the method is provided, including the accuracy of the segmentation and the assumptions made about the statistical distribution of the data. Voxel-based morphometry of MRI data includes spatially normalizing all the images to the same stereotactic space, extracting the gray matter from the normalized images, smoothing, and finally performing a statistical analysis to localize, and make inferences about, group differences. The output from the method is a statistical parametric map showing regions where gray matter concentration differs significantly between groups [13].

**The sequence of "preprocessing → quality check → smoothing → statistical analysis" remains the same for every VBM analysis**

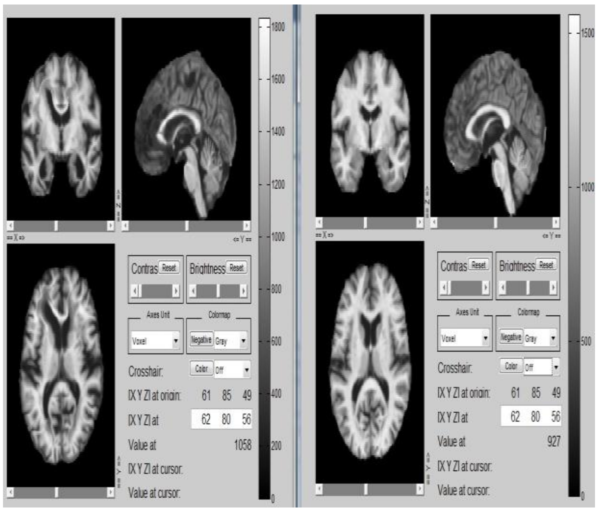
1. T1 images are normalized to a template space and segmented into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF). The preprocessing parameters can be adjusted via the module "Estimate and write".

2. After the preprocessing is finished, a quality check is highly recommended. This can be achieved via the modules "Display one slice for all images" and "Check sample homogeneity using covariance". Both options are located under "VBM8 →Check data quality".

3. Before entering the GM images into a statistical model, image data need to be smoothed. Of note, this step is not

implemented into the VBM8 Toolbox but achieved via the standard SPM module “Smooth”.

The photos before preprocessing its dimension was (176 X 208 X 176) and after VBM and preprocessing steps its dimension reduced to (121 X 145 X 121) voxels size. As shown in Figure 6, which shows the two different photos for demented and nondemented subjects with AD.



Axes Unit: Voxel
Colormap: Gray
Crosshair: Off
[X Y Z] at origin: 61 85 49
[X Y Z] at: 62 80 56
Value at 1058
[X Y Z] at cursor:
Value at cursor:

Axes Unit: Voxel
Colormap: Gray
Crosshair: Off
[X Y Z] at origin: 61 85 49
[X Y Z] at: 62 80 56
Value at 927
[X Y Z] at cursor:
Value at cursor:

**(a) Demented and mild with AD subject**

**(b) Nondemented with AD subject**

**Fig 6: Demented and nondemented subjects with AD after preprocessing**

## 3.2 Dimension Reduction

After the preprocessing each image has now 2122945 Voxels, we need to reduce these number of voxels, so we need to make dimension reduction of each image.

# 4. The Proposed Algorithm

A limitation of PCA and LDA is that when dealing with image data, the image matrices must be first transformed into vectors that are usually of very high dimensionality. This causes expensive computational cost and sometimes the singularity problem.

In the proposed approach, there are 120 different images (49 demented and 71 nondemented) and each of these images has a voxel size of (121 X 145 X 121) which equal to 2122945 voxel. This is very high dimension, so this need to reduce the dimension of each image by neglecting some voxels that is the same in all image and selecting and keeping the other voxels. Proposed approach for this high dimensionality reduction depends on removing the same voxels in all images, which will increase the accuracy and keeping the different voxels. After this step we found that the dimension of each image became equal to 690432 voxels.

## 4.1 Proposed Algorithm steps

**Step1**: Reducing the dimensionality of each images from 2122945 voxels to 690432 voxels by selecting the voxels that have the intensity level different at all image and removing the voxels that have the same intensity level at all images,

**Step2:** partitioning the subjects into two classes, the first includes the images of demented subjects and the other contains the images of nondemented subjects.

**Step3:** Maximize class seperability, by calculating the mean of each class as in Figure 7 that $\mu_1$ is the mean of first class and $\mu_2$ is the mean of the second class.

$$\mu_1 = \frac{1}{n} * \sum_{i=1}^{n} x_i \qquad (5)$$

Where (*n*) is number of images in first class.

$$\mu_2 = \frac{1}{m} * \sum_{j=1}^{m} x_j \qquad (6)$$

Where (*m*) is number of images in second class.

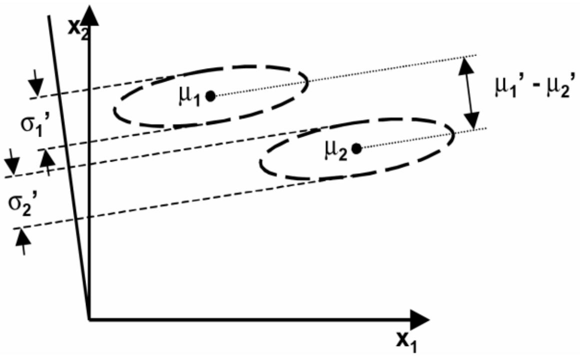


**Fig 7: mean of each classes**

**Step 4:** Calculating standard deviation for each class

$$\sigma_1 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_1)}{n-1}} \qquad (7)$$

$$\sigma_2 = \sqrt{\frac{\sum_{i=1}^{m}(y_j - \mu_2)}{m-1}} \qquad (8)$$

**Step 5:** Maximizing the difference between the means of the two classes by subtract the means and dividing on the multiplication of the standard deviation of two classes

$$w = \frac{|\mu_1 - \mu_2|}{\sigma_1 * \sigma_2} \qquad (9)$$

Where $\mu_1$ and $\mu_2$ are the means of first and second class respectively and $\sigma_1$ and $\sigma_2$ are the standard deviation of first and second class respectively.

**Step 6:** Sorting the result of w in descending. And selecting p numbers of highest w. in selecting we did two selection experiments. The first we choose the first number of w that make the accuracy equal 90%, 80 %, and so on. The second we choose the first number of variables and calculate the accuracy for them. This all is shown in result part.

**Step 7:** Passing these number of w to SVM to make training and testing and calculate the accuracy as shown in next section.

## 4.2 Cross validation

To validate the performance of a learning algorithm, we sometimes use cross validation to provide an empirical measure of the generalization performance. Cross validation is not only used to rate the performance of an algorithm, it is also often applied to tune parameters, such as the regularization parameter in ridge regression. Cross validation techniques involve splitting the full dataset into a training set and a test set, repetitively. At each repetition, the algorithm is trained using the training set, and the trained model is applied to the test set. The average error over all the iterations, between the predicted outcome of the test set and the real target outcome, gives the test error. The most common method is called K-fold cross-validation. The procedure works by partition the dataset into K equal size subsets. For each validation, K-1 subsets (folds) are trained and the remaining fold is used for testing. The procedure will loop K times. At each iteration, a different subset will be chosen as the new testing set. This ensures all the samples will be including in the testing set at least once. If K equals the size of the training set, then at each validation run, only one sample will be left out, hence it is called the leave-one-out cross-validation (loocv) [18].

In the proposed algorithm, the data set is about 120 samples and a cross-validation is done using 5-fold by randomly choosing and making the 5 folds, then training with 4 folds and test with the fifth. The results is shown in result section.

## 4.3 Support Vector Classification

After performing cross-validation, the predicted labels (for classification) was obtained. Support Vector Machine is used as a classifier as it has gained in popularity in recent years because of its superior performance in practical applications, especially in the field of bioinformatics. A two-class support vector machine (SVM) classifier aims to do a hyperplane that maximizes the margin, which is the distance between the closest points on either side of the boundary. These points are known as the support vectors, and their role in the construction of a maximum-margin hyperplane is illustrated in Figure 8. The original SVM algorithm was a linear classifier, but there have since been modifications to deal with data that are not linearly separable. A soft-margin formulation, which allows for mislabeled data as well as a way to use the kernel trick to create nonlinear classifiers. These three formulations are described in further detail in the following subsections [19].

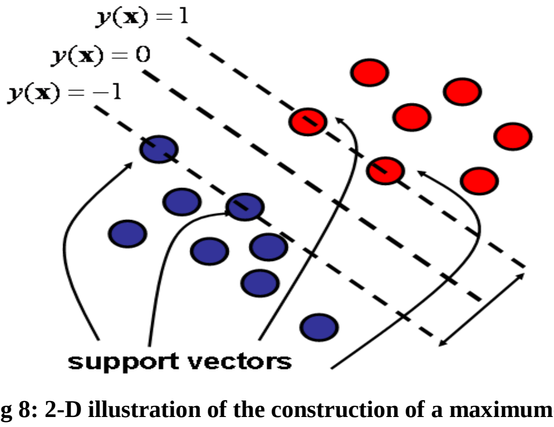


**Fig 8: 2-D illustration of the construction of a maximum-margin hyperplane. This decision surface maximizes the distance between the support vectors, indicated by the arrow.**

### 4.3.1 Linear SVM

The decision surface of a linear SVM classifier is described by y(x) = $w^T * x_i$ - b = 0, as for the Fisher linear discriminant function classifier. The feature weight vector w and threshold b are then chosen such that the margin, or distance between the support vectors, is maximized.

As illustrated in Figure 8, the support vectors lie on two parallel hyperplanes described by y(x) = 1 and y(x) = -1, such that the distance between them is 2/||w||. The maximization of the margin can therefore be expressed as the constrained optimization [19]

$$min_{w,b}\ \frac{1}{2}\ w^T w\ subject\ to\ t_i(w^T x_i - b) \geq 1 \qquad (10)$$

Where the constraint ensures that no feature vectors fall within the margin. By using Lagrange multipliers, this may be re-expressed as the unconstrained optimization [19]

$$min_{w,b} max_{\alpha} \left\{ \{\ \frac{1}{2}\ w^T w - \sum_{i=1}^{N} \alpha_i [t_i(\ w^T x_i - b) - 1] \right)$$

$$subject\ to\quad \alpha_i \geq 0 \qquad (11)$$

From which an expression for the feature weight vector w can be derived in terms of a linear combination of the feature vectors, [19]

$$w = \sum_{i=1}^{N} \alpha_i\, t_i x_i \qquad (12)$$

The decision surface is thus expressed in terms of the support vectors, since only their corresponding $\alpha_i$ are non-zero. A robust solution for the threshold b may then be found by averaging over the $N_{sv}$ support vectors, [19]

$$b = \frac{1}{N_{sv}} \sum_{i=1}^{N_{sv}} (w^T x_i - t_i) \qquad (13)$$

The primal form of the Lagrangian L (w, b, α) may be equivalently written in dual form by substituting the above expression for w. The dual form, [19]

$$max_{\alpha} L(\alpha) = max_{\alpha} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j t_i t_j x_i^T x_j \right\}$$

$$subject\ to\ \alpha_i \geq 0\ and\ \sum_{i=1}^{N} \alpha_i t_i = 0 \qquad (14)$$

Expresses the optimization criterion in terms of inner products of the feature vectors. This is an important property for the creation of nonlinear SVM classifiers [19].

### 4.3.2 Soft-margin SVM

The soft-margin SVM formulation may be applied in cases where no linear hyperplane exists which can separate the data. Slack variables E are introduced, which measure the degree of misclassification of the feature vectors. The optimization becomes a trade-off between maximizing the margin and minimizing the degree of misclassification. This trade-off is controlled by the penalty parameter C, such that the constrained optimization may be expressed as [19]

$$min_{w,\xi,b} \left\{ \frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i \right\}$$

$$subject\ to\ t_i(w^T x_i - b) \geq 1 - \xi_i\ and\ \xi_i \geq 0 \qquad (15)$$

By using Lagrange multipliers, the problem may be re-written as the unconstrained optimization [19]

$$min_{w,\xi,b} max_{\alpha,\beta} \left\{ \frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i [t_i(w^T x_i - b) - 1 + \xi_i] - \sum_{i=1}^{N} \beta_i \xi_i \right\} \quad subject\ to\ \alpha_i, \beta_i \geq 0 \qquad (16)$$

Which may be written in dual form as [19]

$$max_\alpha L(\alpha) = max_\alpha\{\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{ij}\alpha_i\alpha_j t_i t_j x_i^T x_j\}$$

$$subject\ to\ 0 \le \alpha_i \le C\ and\ \sum_{i=1}^{N}\alpha_i t_i = 0 \qquad (17)$$

The only change from the linear SVM optimization is the upper bound on the $\alpha_i$ [19].

### 4.3.3 Nonlinear SVM

In cases where the data are not linearly separable in the input feature space, a nonlinear function ϕ(x) may be used to map each feature vector into a higher-dimensional space. As illustrated in Figure 9, the data are separated by a linear hyperplane in this new space.

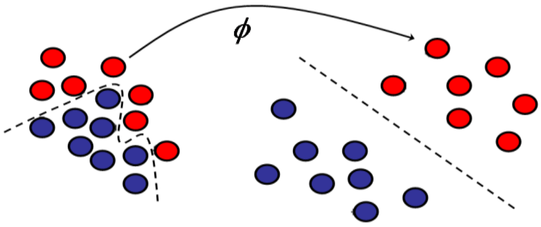


**Fig 9: A nonlinear boundary in the input feature space becomes a linear hyperplane in a higher-dimensional space to which feature vectors are mapped using the nonlinear function ϕ.**

The linear SVM algorithm may then be solved in the transformed feature space by optimizing the dual form Lagrangian [19]

$$L(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{ij}\alpha_i\alpha_j t_i t_j \phi(x_i)^T\phi(x_j) \qquad (18)$$

The optimization criterion is thus expressed in terms of inner products of the transformed feature vectors. By choosing the nonlinear mapping ϕ such that these inner products can be expressed in terms of a kernel function $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$, it is not necessary to explicitly perform the mapping. The optimization problem may therefore be solved even in very high dimensional spaces. The most commonly used kernel is the Gaussian radial basis function, given by

$k(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$, where  > 0 describes the width [19].

**In the proposed algorithm a default Linear Support Vector Classifier is used.**

## 5. Experimental Results

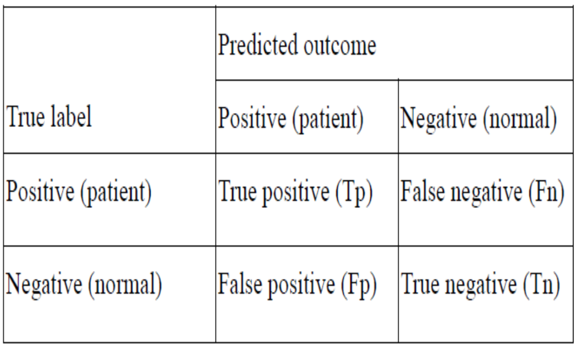
For evaluating classification results, the simplest measurements would be the classification accuracy rate, which is calculated from the number of correctly predicted samples divided by the total number of predicted samples. Often, a single measurement is not sufficient, especially in the cases of disease diagnosis, when the costs of classifying patients into normal and the reverse are not the same. To test the results we used true positive, true negative, false positive and false negative as shown in Figure 10.

True Positive (TP): positive samples correctly classified as positive.

False Positive (FP): positive samples incorrectly classified as negative.

True Negative (TN): negative samples correctly classified as negative.

False Negative (FN): negative samples incorrectly classified as positive.



| True label | Predicted outcome | |
|---|---|---|
| | Positive (patient) | Negative (normal) |
| Positive (patient) | True positive (Tp) | False negative (Fn) |
| Negative (normal) | False positive (Fp) | True negative (Tn) |

**Fig 10: TP, TN, FP, and FN**

The Sensitivity (SEN) of the classifier is the number of true positives (TP) divided by the total number of real positives. In the example, it will be the number of patients. The Specificity (SPE) of the classifier is the number of true negatives (TN) divided by the total number of real negatives (controls). The Accuracy (ACC) number of true positive (TP) plus number of true negative (TN) divided by the sum of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The Accuracy is calculated when number of positive samples equal to number of negative samples, but when positive and negative samples are not equal, then we need to calculate Matthews correlation coefficient (MCC) which gives an accurate description to the Accuracy. As shown in next equations

$$SEN = \ TP/((TP+FN)) \qquad (19)$$

$$SPE = \ TN/((TN+FP)) \qquad (20)$$

$$ACC = \ ((TN+TP))/((TP+TN+FP+FN)) \qquad (21)$$

$$MCC = \ (((TP*TN)-(FP*FN)))/\surd(((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))) \qquad (22)$$

We calculated these fthe parameters for PCA, LDA and the proposal. Then we apply these fthe parameters on two experiments. First, we get the first 10000, 9000, 8000, 7000, and 6000 features, then we apply these features on SVML with PCA, SVML with LDA, and the proposed algorithm with SVML on the same conditions. As shown in tables 3, 4, 5 and 6. And in figures 11, 12, 13 and 14.

In table 3 and figure 11, the Sensitivity was measured for SVML+PCA, SVML+LDA, and SVML+Proposed. Table 4 and figure 12 show the Specificity for SVML+PCA, SVML+LDA, and SVML+Proposed. Table 5 and figure 13 present the Accuracy for SVML+PCA, SVML+LDA, and SVML+Proposed. And table 6 and figure 14 give the Matthews's correlation coefficient for SVML+PCA, SVML+LDA, and SVML+Proposed.

**Table 3. Sensitivity (SEN) for SVML+PCA, SVML+LDA, and SVML+Proposed**

| Algorithm | No. of Features | | | | |
|---|---|---|---|---|---|
| | 10000 | 9000 | 8000 | 7000 | 6000 |
| SVML+PCA | 60.9% | 59.6% | 58.7% | 57.8% | 56.8% |
| SVML+LDA | 100% | 100% | 100% | 100% | 100% |
| SVML+Proposed | 100% | 100% | 100% | 100% | 94.2% |

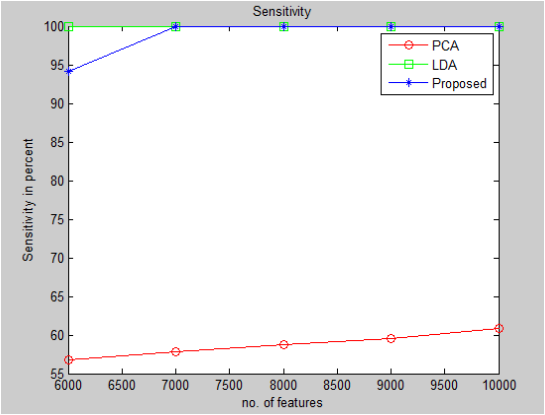


**Fig 11: the Sensitivity (SEN) for SVML+PCA, SVML+LDA, and SVML+Proposed**

**Table 4. Specificity (SPE) for SVML+PCA, SVML+LDA,**

| Algorithm | No. of Features | | | | |
|---|---|---|---|---|---|
| | 10000 | 9000 | 8000 | 7000 | 6000 |
| SVML+PCA | 71.6% | 71.2% | 70.3% | 69.3% | 68.4% |
| SVML+LDA | 93.4% | 93.4% | 93.4% | 94.7% | 95.9% |
| SVML+Proposed | 93.4% | 97.3% | 100% | 100% | 100% |

**and SVML+Proposed**

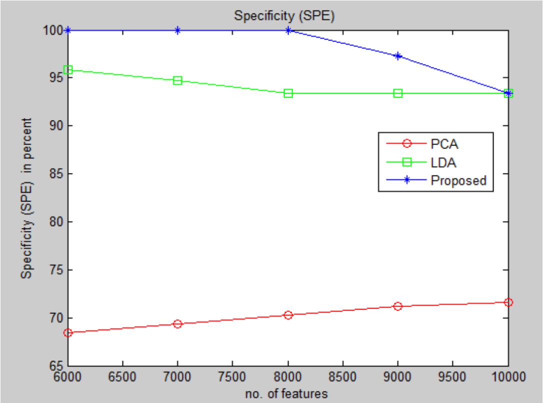


**Fig 12: Specificity (SPE) for SVML+PCA, SVML+LDA, and SVML+Proposed**

**Table 5. Accuracy (ACC) for SVML+PCA, SVML+LDA, and SVML+Proposed**

| Algorithm | No. of Features | | | | |
|---|---|---|---|---|---|
| | 10000 | 9000 | 8000 | 7000 | 6000 |
| SVML+PCA | 67.5% | 66.7% | 65.8% | 65% | 64.2% |
| SVML+LDA | 95.8% | 95.8% | 95.8% | 96.7% | 97.5% |
| SVML+Proposed | 95.8% | 98.3% | 100% | 100% | 97.5% |

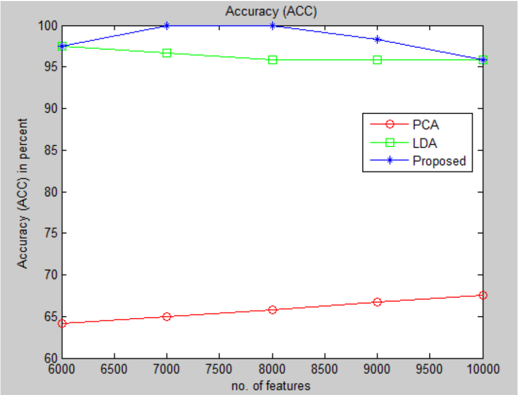


**Fig 13: Accuracy (ACC) for SVML+PCA, SVML+LDA, and SVML+Proposed**

**Table 6. Matthews's correlation coefficient (MCC) for SVML+PCA, SVML+LDA, and SVML+Proposed**

| Algorithm | No. of Features | | | | |
|---|---|---|---|---|---|
| | 10000 | 9000 | 8000 | 7000 | 6000 |
| SVML+PCA | 32.1% | 30.6% | 28.7% | 26.7% | 24.7% |
| SVML+LDA | 91.6% | 91.6% | 91.6% | 93.2% | 94.9% |
| SVML+Proposed | 91.6% | 96.6% | 100% | 100% | 95% |

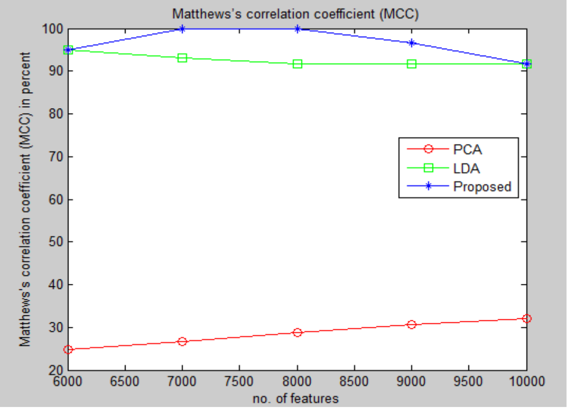


**Fig 14: Matthews's correlation coefficient (MCC) for SVML+PCA, SVML+LDA, and SVML+Proposed**

In the second experiment, we get 90%, 80%, 70%, 60% and 50% of the features, then we apply these features on SVML with PCA, SVML with LDA, and the proposed algorithm with SVML on the same conditions. As shown in tables 7, 8, 9 and 10, and figures 15, 16, 17, and 18. In table 7 and figure 15, the Sensitivity was measured for SVML+PCA, SVML+LDA, and SVML+Proposed. Table 8 and figure 16 show the Specificity for SVML+PCA, SVML+LDA, and

SVML+Proposed. Table 9 and figure 17 present the Accuracy for SVML+PCA, SVML+LDA, and SVML+Proposed. And table 10 and figure 18 give the Matthews's correlation coefficient for SVML+PCA, SVML+LDA, and SVML+Proposed.

**Table 7.Sensitivity (SEN) for SVML+PCA, SVML+LDA, and SVML+Proposed**

| Algorithm | No. of Features in percent | | | | |
|---|---|---|---|---|---|
| | 90% | 80% | 70% | 60% | 50% |
| SVML+PCA | 66.7% | 70% | 70% | 69% | 70% |
| SVML+LDA | 83% | 68.3% | 87.2% | 85.7% | 91.3% |
| SVML+Proposed | 100% | 100% | 89.1% | 94.2% | 100% |

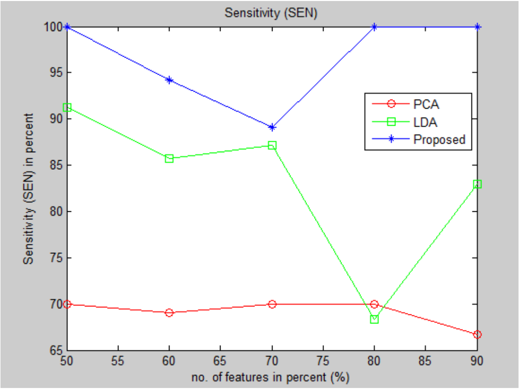


**Fig 15: the Sensitivity (SEN) for SVML+PCA, SVML+LDA, and SVML+Proposed**

**Table 8. Specificity (SPE) for SVML+PCA, SVML+LDA, and SVML+Proposed**

| Algorithm | No. of Features in percent | | | | |
|---|---|---|---|---|---|
| | 90% | 80% | 70% | 60% | 50% |
| SVML+PCA | 73.1% | 73.8% | 73.8% | 74.4% | 73.8% |
| SVML+LDA | 86.3% | 89.5% | 81.5% | 90.1% | 90.5% |
| SVML+Proposed | 100% | 100% | 100% | 100% | 98.6% |

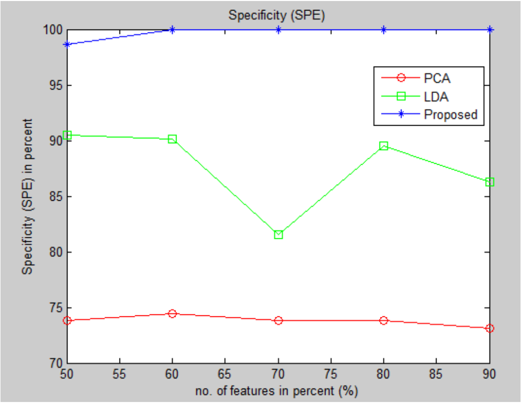


**Fig 16: Specificity (SPE) for SVML+PCA, SVML+LDA, and SVML+Proposed**

**Table 9. Accuracy (ACC) for SVML+PCA, SVML+LDA, and SVML+Proposed**

| Algorithm | No. of Features in percent | | | | |
|---|---|---|---|---|---|
| | 90% | 80% | 70% | 60% | 50% |
| SVML+PCA | 70.8% | 72.5% | 72.5% | 72.5% | 72.5% |
| SVML+LDA | 85% | 78.3% | 83.3% | 88.3% | 90.8% |
| SVML+Proposed | 100% | 100% | 95% | 97.5% | 99.2% |

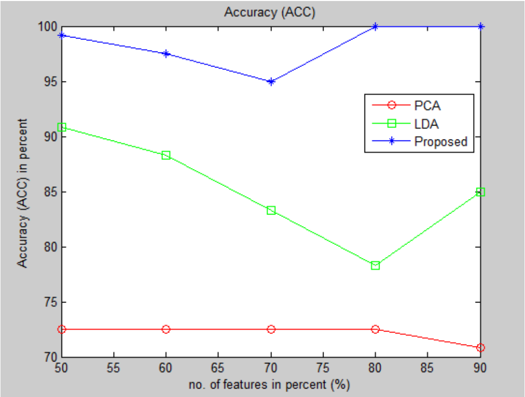


**Fig 17: Accuracy (ACC) for SVML+PCA, SVML+LDA, and SVML+Proposed**

**Table 10. Matthews's correlation coefficient (MCC) for SVML+PCA, SVML+LDA, and SVML+Proposed**

| Algorithm | No. of Features in percent | | | | |
|---|---|---|---|---|---|
| | 90% | 80% | 70% | 60% | 50% |
| SVML+PCA | 38.6% | 42% | 42% | 42.1% | 42% |
| SVML+LDA | 68.8% | 58.6% | 65.4% | 75.9% | 81% |
| SVML+Proposed | 100% | 100% | 90.3% | 95% | 98.3% |

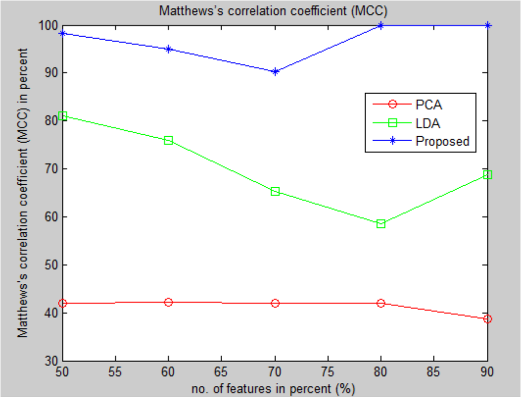


**Fig 18: Matthews's correlation coefficient (MCC) for SVML+PCA, SVML+LDA, and SVML+Proposed**

## 6. Result Discussion

In the first experiment, we get the first number of features equal 10000, 9000, 8000, 7000, and 6000, then we first calculated the Sensitivity (SEN) for PCA, LDA, and the proposed algorithm. The PCA gives sensitivity equal 57.8% at 7000 features but LDA and the proposed algorithm give sensitivity equal to 100% as shown previously in table 3. Second, The Specificity (SPE) was also calculated for PCA, LDA, and the proposed algorithm. The PCA gives Specificity equal 69.3% at 7000 features, LDA gives specificity equal to 94.7% and the proposed algorithm give specificity equal to 100% as shown previously in table 4. Third, the Accuracy (ACC) was calculated for PCA, LDA, and the proposed algorithm. The PCA gives accuracy equal 65% at 7000 features, LDA gives accuracy equal to 96.7% and the proposed algorithm give accuracy equal to 100% as shown previously in table 5. Last, the Matthews's correlation coefficient (MCC) was calculated for PCA, LDA, and the proposed algorithm. The PCA gives MCC equal 26.7% at 7000 features, LDA gives MCC equal to 93.2% and the proposed algorithm give MCC equal to 100% as shown previously in table 6.

In the second experiment, we get the first 90%, 80%, 70%, 60% and 50% of the features, then we first calculated the Sensitivity (SEN) for PCA, LDA, and the proposed algorithm. The PCA gives sensitivity equal 70% at 50% of the features but LDA gives sensitivity equal to 91.3 and the proposed algorithm gives sensitivity equal to 100% as shown previously in table 7. Second, The Specificity (SPE) was also calculated for PCA, LDA, and the proposed algorithm. The PCA gives Specificity equal 73.8% at 50% of the features, LDA gives specificity equal to 90.5% and the proposed algorithm give specificity equal to 100% as shown previously in table 8. Third, the Accuracy (ACC) was calculated for PCA, LDA, and the proposed algorithm. The PCA gives accuracy equal 72.5% at 50% of the features, LDA gives accuracy equal to 90.8% and the proposed algorithm give accuracy equal to 99.2% as shown previously in table 9. Last, the Matthews's correlation coefficient (MCC) was calculated for PCA, LDA, and the proposed algorithm. The PCA gives MCC equal 42% at 50% of the features, LDA gives MCC equal to 81% and the proposed algorithm give MCC equal to 98.3% as shown previously in table 10.

After the experiments we found that the proposed Algorithm gives Accuracy of 100% with little number of voxels reach to 6610 voxels. We found that the difference between the intensity level of highest and lowest pixels equal to 42. This means that it is impossible to detect that this subject is demented or non-demented by human visual.

## 7. CONCLUSION

In this work we have studied feature extraction processes based on VBM analysis, to classify MRI volumes of AD patients and normal subjects. We have analyzed the data set using the SPM with the VBM to normalize the images, then we studied several approaches for the automatic classification of Alzheimer's disease. After that we talk about feature selection and extraction techniques and different techniques of Support Vector Machine. Then we knew the different techniques of each one of them and we compared the proposed algorithm with PCA and LDA. After that we apply PCA, LDA, and the proposal to Linear Support Vector Classifier, we found that the accuracy reached to 100% for the proposed algorithm.

## 8. ACKNOWLEDGMENTS

The thanks to the experts who have contributed towards development of the template.

## 9. REFERENCES

[1] P. Morgado, "Automated Diagnosis of Alzheimer's Disease using PET Images", MSc thesis at Electrical and Computer Engineering Dep., Higher technical institute, Technical University of Lisbon, September 2012.

[2] C. P. Ferri, R. Sousa, E. Albanense, W. s. Ribeiro, and M. Honyashiki, "World Alzheimer Report 2009," 2009.

[3] A. Wimo and M. Prince, "World Alzheimer Report 2010: The global economic impact of dementia," September 2010.

[4] A. Association, "2012 Alzheimer's disease facts and figures," Alzheimer's and Dementia: The Jthenal of the Alzheimer's Association, vol. 8, no. 2, pp. 131–168, 2012.

[5] P. Padilla, M. López, J. M. Górriz, J. Ramirez, D. Salas-Gonzalez, and I. Álvarez, "NMF-SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's disease," Medical Imaging, IEEE Transactions on, vol. 31, no. 2, pp. 207–216, 2012.

[6] F.J. Martinez-Murcia , J.M. Gorriz , J. Ramirez , C.G. Puntonet , D. Salas-Gonzalez or the Alzheimer's Disease Neuroimaging Initiative, "Computer Aided Diagnosis tool for Alzheimer's Disease based on Mann–Whitney–Wilcoxon U-Test", Expert Systems with Applications 39 (2012) 9676–9685

[7] R. Chaves, J. Ramirez a, J.M. Gorriz, C.G. Puntonet , Alzheimer's Disease Neuroimaging Initiative, "Association rule-based feature selection method for Alzheimer's disease diagnosis", Expert Systems with Applications , 2012.

[8] Ashburner J, Friston KJ. Voxel-based morphometry—the methods. Neuroimage 2000;11:805–21.

[9] Adam M. Brickman , Christian Habeck, Eric Zarahn, Joseph Flynn, Yaakov Stern, "Structural MRI covariance patterns associated with normal aging and neuropsychological functioning", Neurobiology of Aging, 2006

[10] VBM8: http://dbm.neuro.uni-jena.de/vbm/

[11] SPM8: http://www.fil.ion.ucl.ac.uk/spm/

[12] OASIS Data set: http://www.oasis-brains.org/

[13] John Ashburner and Karl J. Friston, "Voxel-Based Morphometry—The Methods", NeuroImage 11, 805–821 (2000).

[14] P. Cunningham, "Dimension Reduction", University College Dublin, Technical Report UCD-CSI-2007-7, 2007

[15] Isabelle Guyon , Andr´e Elisseeff, "An Introduction to Feature Extraction", Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.

[16] Lindsay I. Smith," A tutorial on Principal Components Analysis", University of Otago, New Zealand 2002.

[17] Shih, Frank Y,"Image processing and pattern recognition: fundamentals and techniques.", IEEE,2010.

[18] Chia-Yueh C. CHU, "Pattern recognition and machine learning for magnetic resonance images with kernel methods", thesis submitted for the degree of Doctor of Philosophy, University College London, 2009.

[19] Katherine R. Gray, "Machine learning for image-based classification of Alzheimer's disease", thesis submitted for the degree of Doctor of Philosophy, Department of Computing, Imperial College London, 2012.

[20] Y. Xia, L. Wen, S. Eberl, M. Fulham, and D. Feng, "Genetic algorithm-based PCA eigenvector selection and weighting for automated identification of dementia using FDG-PET imaging," in Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, pp. 4812–4815, 2008.

[21] I. Illán, J. Górriz, J. Ramírez, D. Salas-Gonzalez, M. López, F. Segovia, R. Chaves, M. Gómez-Rio, and C. Puntonet, "18F-FDG PET imaging analysis for computer aided Alzheimer's diagnosis," Information Sciences, vol. 181, no. 4, pp. 903–916, 2011.

[22] M. López, J. Ramírez, J. Górriz, D. Salas-Gonzalez, I. Álvarez, F. Segovia, and R. Chaves, "Multivariate approaches for Alzheimer's disease diagnosis using Bayesian classifiers," in Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE, pp. 3190–3193, 2009.