Journal of Theoretical and Applied Information Technology
15th February 2021. Vol.99. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# FEATURE EXTRACTION OF ALZHEIMER'S DISEASE CLASSIFICATION BASED ON PRINCIPAL COMPONENT AND RANDOM SUBSPACE DISCRIMINANT ANALYSIS

**AOW YONG LI YEW[1], MOHD SHAFRY MOHD RAHIM[2],**

ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

[1,2] School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

E-mail: [1]ylyaow2@live.utm.my, [2]shafry@utm.my

**ABSTRACT**

Alzheimer's disease (AD) is one of the diseases which brings great influences on the lives of the people. AD classification can serve as a supportive tool to help the doctor to analyze the brain images. One of the important steps in AD classification is feature extraction. Among the feature extraction techniques, Principal Component Analysis (PCA) is a widely used machine learning approach. Nevertheless, it is hard to decide the number of dimensions to be extracted after the transformation. The accuracy of the classification can be greatly affected by the number of dimensions to be chosen. Therefore, this paper has developed a feature extraction method based on principal component and random subspace discriminant analysis (PCRSDA) to extract and select the features. The selection of the number of dimensions was determined by 10-fold cross validation where the features were selected randomly without replacement. The dataset in this paper was collected from Alzheimer's Disease Neuroimaging Initiative (ADNI) database across four time points. The classification results were 81%, 84%, 87% and 87% at time point of 24 months before stable diagnosis, 18 months before stable diagnosis, 12 months before stable diagnosis and at the stable diagnosis time point, respectively.

**Keywords:** *Alzheimer's Disease, Principal Component Analysis, Feature Extraction, Classification, Machine Learning*

## 1. INTRODUCTION

Computer-aided diagnosis (CAD) is one of the contributions of computer vision. It helps the doctor to predict the disease and observe the progression of the disease through medical imaging. Magnetic resonance imaging (MRI) is one of the commonly used medical imaging due to its non-invasive procedure. In Alzheimer's Disease (AD) diagnosis, structural MRI is treated as a supportive feature [1]. Doctors can infer the possibility of the development of the disease through observing the atrophy of the brain from the magnetic resonance images. The atrophy of medial temporal lobe (MTL) is the main biomarker in AD diagnosis. However, the distribution of affected areas is extended to outside of MTL. Therefore, it attracts many researchers to identify the significant features for AD diagnosis. The process of identifying the significant features in CAD is called feature extraction or feature selection. This is a crucial step in machine learning approach before classification.

Feature extraction and feature selection serve with the same purposes, which are to identify the significant features while reducing the number of dimensions for classification. The difference of both processes is feature extraction involves data transformation while feature selection identifies the useful features from raw data [2], [3]. Despite this, the techniques can be used separately, or it can be conducted at the same time. In AD classification, many techniques were adopted such as Principal Component Analysis (PCA) [4], Partial Least Squares (PLS) [5], ReliefF Algorithm [6], and the combination of T-test feature ranking and fisher criterion [7].

Among the techniques, PCA is the most used technique because of its simplicity and outstanding performance. PCA is an unsupervised approach which transforms the data by maximizing the variance of the data. It decomposes the data to uncorrelated data, which is called principal

components. The first principal component has the highest variance, and it presents most of the data. The easiest way to determine the number of principal components to be extracted is to set a threshold for the percentage of total variance explained by each principal component. Besides, we also can use the scree plot to find the cumulative sum of the variance. The cumulative sum of the variance drops dramatically will be the number of dimensions to be extracted. There are problems with the selection methods. The first method requires a manually set for the threshold. The second method is not feasible for AD classification due to the drastic change of cumulative sum of the variance only happened at first few principal components, but it is not sufficient for achieving higher accuracy in the AD classification. Therefore, the selection of the number of dimensions or principal components remain an issue for AD classification.

Apart from the methods have been mentioned above, intrinsic dimension estimation methods can be the problems solver. Intrinsic dimension refers to the number of dimensions needed to interpret the data before data transformation. We applied this term in the following sections to represent the number of dimensions needed to be extracted after PCA. The well-known intrinsic dimension estimation methods are Maximum Likelihood Estimator (MLE), Eigenvalue-based Estimation (EigValue) and Correlation Dimension Estimator (CorrDim) [8]–[10]. The more recent method is Dimensionality from Angle and Norm Concentration (DANCo) [11], [12]. We have examined the ability of these methods by conducting a comparison with the proposed method in this paper.

The main contribution of this paper is to develop a feature extraction approach for AD classification by solving the feature selection issue of PCA. The proposed methods take advantages on the strength of Random Subspace Method (RSM) and Linear Discriminant Analysis (LDA) to select the number of principal components. RSM [13], [14]reduces the bias of each learner through random sampling the features [14]. LDA works well in deciding the set of variables which is effective in predicting the class membership [15]. By combining both methods with PCA, the proposed method becomes less sensitivity to the number of observations in selecting the significant features. Therefore, the model can generalize to test set better. This can be proved through the results and discussion sections. The sections of this paper are organized as follows: Section 2 presents the related work, Section 3 describes the dataset and methodology, Section 4 discusses the results and Section 5 gives the conclusion of this paper.

## 2. RELATED WORKS

There are many researchers have been conducted in AD classification. PCA was the widely used methods in feature extraction for longitudinal study and cross-sectional study [16]–[19]. In [17], [19], PCA was applied to find the significant cortical features for AD classification. The authors conducted a two-sample T-test to examine the discriminative power of individual vertex. In [18], the authors used PCA to extract the volumetric features for longitudinal study. It took the total variance explained of the principal components to decide the intrinsic dimension. The best classification result was obtained by using all principal components with radial basis function SVM.

In [16], PCA was applied to find the significant shape features for AD classification. The authors used a bagging approach to select the features. The paper also mentioned the wrapper methods which is similar to the proposed method. The bagging approach resamples the data and feeds it to multiple classifiers. The result is decided through majority voting. The wrapper approach conducts the feature selection with the classifier used in the classification process. On the other hand, the proposed method is the combination of wrapper and bagging approach. The proposed method randomly selects the features instead of resampling the data. The classifier used in feature selection process is different with the classifier used in the classification process. However, it is not possible to compare both techniques because the features used in both studies were different. In [16], the features was shape of the hippocampus while the features in this study was volumetry of the whole brain gray matter.

On the other hand, the intrinsic dimension estimation such as MLE finds the intrinsic dimension through maximizing the likelihood to the distances between close neighbours [8]. The distances between close neighbours are computed through k nearest neighbour (KNN). In general, the researchers suggested k equal to the square root of the training set size [20]. Therefore, we adopted this concept when we evaluated the intrinsic dimension estimation methods. EigValue method determines the intrinsic dimension by comparing the eigenvalues with the threshold [10]. The value of the threshold should not be too high because the purpose

for dimensionality reduction is extracting significant features while minimizing the loss of information. CorrDim finds the distance of the nearest neighbourhood through KNN [9]. By using the median value and maximum value of KNN, CorrDim estimates the intrinsic dimension by calculating the slope of the curve. DANCo is specified design for the high dimensionality data through considering the concentration of distance and the angle effects in high dimensionality data [11], [12]. DANCo also applies KNN to find the nearest neighbours. The concentration and angle effects are calculated through identifying the value of the set of neighbours. There are different pros and cons on each method, this paper investigated the strength of the methods and proposed a new method.

## 3. METHODOLOGY

### 3.1 Data Collection and Preparation

The data were obtained from ADNI database. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. 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). For up-to-date information, see www.adni-info.org.

The data were collected across four time points, which were at the time point of 24 months before stable diagnosis, 18 months before stable diagnosis, 12 months before stable diagnosis and during stable diagnosis time point. The data consisted of 50 healthy control (CN), 50 stable mild cognitive impairment (SMCI), 50 progressive mild cognitive impairment (PMCI) and 50 AD. The subjects were grouped into two classes, which were AD with PMCI and CN with SMCI in this paper. Computational Anatomy Toolbox (CAT12) software were used to apply spatial normalization to transform the same anatomical regions of the brain in each image to the same voxels. The data were segmented to gray matter (GM) and white matter (WM). This paper used GM as the features for AD classification. Figure 1 illustrates the example of original sample and the sample after segmentation. The number of dimensions of the original sample was (256x256x166) while the number of dimensions of the pre-processed sample was (121x145x121). After the preprocessing, the data were partitioned equally to two subsets for training and testing purpose.

### 3.2 Principal Component and Random Subspace Analysis

The proposed method consists of several steps as illustrated in Figure 2. Section 3.2.1 involves computing the uncorrelated data with maximum dimension numbers through PCA. Section 3.2.2 describes the process to find optimal number of dimensions by calculating the error rate, and Section 3.2.3 extracts the principal components.

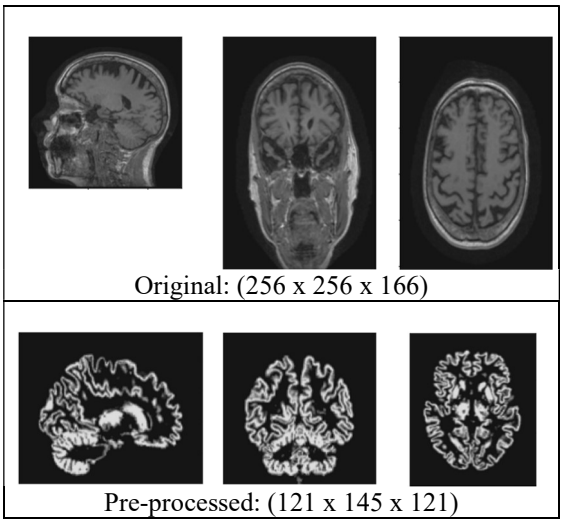


*Figure 1: The Difference between the Original Sample and Pre-processed Sample*

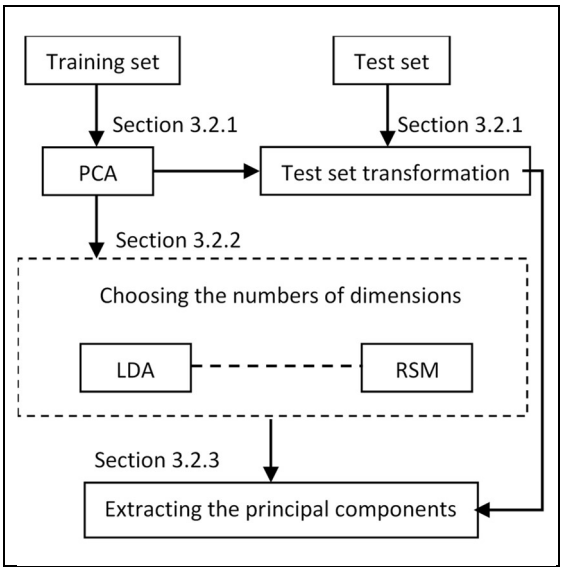


*Figure 2: The Procedures of Feature Extraction*

### 3.2.1 Principal component analysis

The conventional PCA consists of four basic steps to achieve feature extraction. First step is called relationship identification, which identifies the correlation between the data. Second step is called linear transformation or eigen-decomposition. It transforms a set of correlated features into a set of linearly uncorrelated variables. Third step is principal component transformation, which involves transforming the data into principal components. The final step is feature extraction through quantifying the significant values of the principal components. In the final step, the percentage of the total variance explained by each principal component is computed. Then, a threshold is set to determine the number of principal components to be extracted based on the percentage of total variance. Figure 3 demonstrates the brain image with different percentage of total variance. It is hardly to decide the optimal value as the threshold. Hence, this study replaced this step with the process in Section 3.2.2 and the feature extraction was completed in Section 3.2.3. The details of the first three steps used in this study are given as follows:

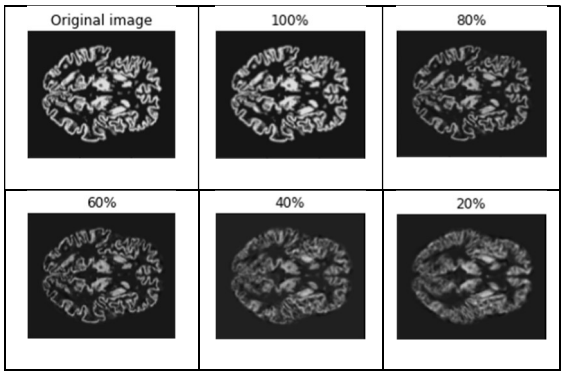


*Figure 3: The Image with Different Percentage of Total Variance*

i. PCA starts with mean centring of training data to reduce multicollinearity. It is essential because multicollinearity makes the interpretation of the coefficients becomes complicated and eventually undermines the significance value of the variables. The mean centred data $X$ has *n-by-p* size, which $n$ is the number of samples and $p$ is the number of variables of each sample. It follows by calculating the covariance matrix of $X$. The covariance matrix $A$ is computed through matrix multiplication of $X$ with $X'$, where $X'$ is the transpose of $X$.
ii. Find the eigenvectors and eigenvalues of the n-by-n covariance matrix, $A$. The eigenvectors and eigenvalues can be calculated through the expansion of the following formula:

$$Av = \lambda v \quad (1)$$

where v are the eigenvectors of $A$ and λ are the eigenvalues or variance of $A$. The maximum dimension of the eigenvectors is (n-1) to gain non-zero eigenvalues. Therefore, the number of principal components, which also considered as the dimension number obtained from PCA is 99 because the training data has 100 samples.
iii. The principal components of matrix $X$ is obtained by multiplying it with $A$. The principal component that explains most of the data will be the first principal component as the eigenvectors are arranged in descending order according to the eigenvalues. At the same time, the test data transformation starts with mean centring by using the mean of training data. Then, the mean centred test data is transformed through multiplying it with $A$.

### 3.2.2 Choosing the number of dimensions

The transformed training data from PCA is the input for the process of choosing the number of dimensions. The model trains LDA using RSM ensemble with various dimension numbers of transformed data. The model is trained with 10-fold cross validation. The d-dimensions of transformed training data is divided to 10 set randomly, where d = {1, 2, 3, …, 99}. 1 out of 10 set is treated as validation set and the rest are used to train the model. Finally, the intrinsic dimension is computed through calculating the classification loss of each number of dimensions. The classification loss is the weighted fraction of misclassified sample. The target is to find out the number of dimensions achieve the lowest classification loss.

### Linear Discriminant Analysis

LDA is a linear classifier which intends to find the axes that maximize the separation of known classes [21]. There are two criteria to create the new axes, which are maximizing the distance between the sample mean of each class and minimizing the variation within each class. In order to ease the explanation, some of the notation is clarified as follows. The dataset involves $K$-classes, and there are two classes in this study. Each class has $n_k$ $d$-dimensional samples, where $n_k$ is the number of samples in class $k$. The steps to train the linear discriminant model are described as below:

i. Calculate the sample mean of each class, $\mu_k = \{\mu_{k1}, \mu_{k2}, \dots \mu_{kd}\}$.

$$\mu_k = \frac{1}{n_k}\sum_{x\epsilon k} x \quad (2)$$

ii. Compute within-class and between-class scatter matrices. The within-class scatter matrix is computed as $S_W = S_1 + S_2$ based on Equation 3. The between-class scatter matrix, $S_B$ is computed as Equation 4, where $\mu$ is the mean of all observations regardless the class.

$$S_k = \sum_{x\in k}(x-\mu_k)(x-\mu_k)^T \quad (3)$$

$$S_B = \sum_{k=1}^{K} n_k(\mu_k-\mu)(\mu_k-\mu)^T \quad (4)$$

By multiplying the scatter matrices as $S_W^{-1}S_B$, it fulfils the two criteria mentioned above to find the best axes to separate the two classes.

iii. Find the eigenvectors and eigenvalues for the matrix $S_W^{-1}S_B$. The eigenvectors and eigenvalues are calculated through the expansion of Equation 1 by replacing $A$ to $S_W^{-1}S_B$. The eigenvectors $LD_d$, are arranged according to the eigenvalues of matrix $S_W^{-1}S_B$ in descending order. This refers to the $LD_1$ will be the best axis to separate both classes and the separability level goes down for the rest of the axes.

iv. Project training set and validation set to the new axes according to the class by multiplying it with the eigenvectors.

After the training of linear discriminant model, LDA employs Bayes' Theorem to predict the class label for the validation set through minimizing the expected classification cost according to Equation 5. The notation $\hat{y}$ is the predicted class label, $\hat{P}(k|x)$is the posterior probability of class $k$, given the sample $x$ and $C(y|k)$ is the classification cost of a sample as $y$ when its true class is $k$. The cost is set to 0 if the sample is classified correctly and the cost is set to 1 if the sample is classified incorrectly. The cost matrix is a square matrix of size $K$-by-$K$. Therefore, diagonal of the cost matrix is equal to 0 and the remaining entries are equal to 1.

$$\hat{y} = \underset{y=1,\ldots,K}{arg\,min}\sum_{k=1}^{K}\hat{P}(k|x)C(y|k) \quad (5)$$

The posterior probability is defined as the product of the multivariate normal density $P(x|k)$ and the prior probability $P(k)$, which is normalized with a constant $P(x)$ as the sum over that product of both classes. $P(k)$is the probability of the training samples belong to class $k$. $P(x|k)$ takes the main role in the prediction of class label, which calculate the squared Mahalanobis distance of the sample $x$ from the mean of class $k$. The related equations are written from Equation 6 to Equation 11, where $Cov_k$ is the covariance of class $k$, $Cov_k$ is the determinant of $Cov_k$ and $Cov_k^{-1}$ is the inverse matrix.

$$\hat{P}(k|x) = \frac{P(x|k)P(k)}{P(x)} \quad (6)$$

$$P(x|k) = \frac{1}{((2\pi)^n|Cov_k|)^{1/2}}\,exp(B) \quad (7)$$

$$B = -\frac{1}{2}(x-\mu_k)Cov_k^{-1}(x-\mu_k)^T \quad (8)$$

$$P(k) = \frac{n_k}{N} \quad (9)$$

$$P(x) = \sum_{k=1}^{K}P(x|k)P(k) \quad (10)$$

$$Cov_k = \frac{1}{n_k}\left(\sum_{x\in k}(x-\mu_k)(x-\mu_k)^T\right) \quad (11)$$

**Random Subspace Method**

In this study, the ensemble learning RSM is used to reduce the correlations of the LDA classifiers. RSM trained LDA on randomly chosen subspaces, which means various numbers of dimensions are randomly selected without replacement in each iteration rather than all variables of the samples. This contributes to increase the generalization accuracy by preventing over-focusing on certain variable [22]. RSM is implemented with cross validation to find the intrinsic dimensions for PCA. The steps of the method are reported as following:

i. Initialize the parameters. Let the numbers of dimensions, $m$ = {1,2,3,…$d$}. The loop starts with $m$ = 1. Let the number of LDA learning cycles, $NLCycle = n$.
ii. Divide the training set obtained from PCA to ten sets. Nine out of ten sets are used for training the model and one set is used for validation.
iii. Train each LDA with random set of $m_i$ dimensions without replacement. Repeat this step until $NLCycle$.
iv. Predict on the validation set by computing the average score of the prediction from all learners. The sample is classified to the respective class based on the category obtains highest average score.
v. Calculate the classification error, $err_m$.
vi. Repeat steps 2 and 5 until 10 times since this is a 10-fold cross validation.

vii. Repeat step 6 until $d$ dimensions.

The *NLCycle* is a fixed value to focus on the effect of the numbers of dimensions towards the prediction. The numbers of learners can be from dozens to thousands. The computation cost increases according to the increment of the numbers of learners. Therefore, it is reasonable to set the $NLCycle$ according to the number of samples to balance the computation cost and the discriminative power of the classifier.

### 3.2.3 Extracting the principal components

After calculating the classification error from the training set, the number of dimensions that obtains lowest classification error is the intrinsic dimension for extracting the principal components. Both transformed training set and test set are extracted through eliminating the dimensions after the intrinsic dimension. It is important to exclude test set from the early stage of PCA and during the estimation of intrinsic dimension. This prevents the data leakage problem that causes the model becomes invalid.

### 3.3 Classification

Linear support vector machine (SVM) is adopted in the classification [23]. It is a supervised machine learning approach which builds a trained model that maximizes the hyperplane of two diagnostic groups to predict the new data. The low dimensional training set obtained from previous section is the input for the trained model while the low dimensional test set is used for the prediction. The measurement of the classification is accuracy. The accuracy tells us the ability of the model to differentiate two diagnostic groups correctly [24]. The calculation of accuracy is presented in Equation 12. It is derived from confusion matrix through calculating true positive (TP), true negative (TN), false positive (FP) and false negative (FN) as shown in Table I.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (12)$$

*Table 1: Confusion Matrix*

| | | Predicted Class | |
|---|---|---|---|
| | | AD+PMCI | CN+SMCI |
| Actual Class | AD+PMCI | True Positive | False Negative |
| | CN+SMCI | False Positive | True Negative |

## 4. RESULTS AND DISCUSSION

This section is divided to two parts. Section 4.1 compares the proposed method with intrinsic dimension estimation techniques. Section 4.2 validates the methods through conducting a comparison with different classifiers used in feature extraction.

### 4.1 The Comparison of Proposed Method with Intrinsic Dimension Estimation Methods

In this paper, we included four intrinsic dimension estimation methods, which were MLE, EigValue, CorrDim and DANCo. The techniques were briefly explained in Section 2. However, it is worth to take note that EigValue performs as the aforementioned method to find the number of dimensions. It set a threshold on the percentage of total variance after applying PCA on the raw data.

KNN was used in different intrinsic dimension estimation methods. According to the literature review, k was suggested set according to the square root of training size. Therefore, the k value for MLE was set between 6 to 14, while k value for DANCo and CorrDim was 10. CorrDim found the non-zero elements of the data after computing KNN. The median and maximum value of the output of KNN were used to estimate the intrinsic dimension. On the other hand, the threshold of EigValue was set to 2.5% in this study, which required the eigenvalues interpreted more than 97.5% of the total variance of the data.

Table 2 shows the intrinsic dimensions based on different methods. The estimated intrinsic dimensions were very different based on different approaches, but there are some trends that we can extract from the results. MLE considers more dimensions needed compared to other methods at all time points, which was at least 30 intrinsic dimensions. The second highest intrinsic dimension were computed by using EigValue. But the intrinsic dimensions were significantly lower compared to MLE.

CorrDim obtained the lowest intrinsic dimension, and the results were same for all time points. The symbol "±" for DANCo referred to the approximate value, it was written as approximation because the intrinsic dimensions changed in different rounds of tests and the values obtained were close to the approximate value. The proposed method estimated the intrinsic dimensions were 70 and above at most of the time points except for the

dataset collected at time point of 24 months before stable diagnosis.

Table 3 presents the results of the comparison between proposed method with the intrinsic dimension estimation methods. From the results, we notice that the proposed method outperformed other methods, which achieved 81%, 84%,87% and 87% for the dataset collected at time point of 24 months before stable diagnosis, 18 months before stable diagnosis, 12 months before stable diagnosis and at the stable diagnosis time point, respectively. The results showed that we need more dimensions to represent the data.

*Table 2: The Intrinsic Dimensions based on Different Approaches*

| Methods | Intrinsic Dimensions | | | |
|---|---|---|---|---|
| | 24m | 18m | 12m | Stable |
| MLE | 31 | 38 | 33 | 30 |
| EigValue | 8 | 5 | 6 | 7 |
| CorrDim | 2 | 2 | 2 | 2 |
| DANCo | ±9 | ±9 | ±8 | ±7 |
| Proposed method | 16 | 74 | 70 | 74 |

*Table 3: A Comparison of Proposed Method and Existing Intrinsic Dimension Estimation Methods*

| Methods | 24m | 18m | 12m | Stable |
|---|---|---|---|---|
| MLE | 79 | 81 | 83 | 87 |
| EigValue | 79 | 81 | 84 | 84 |
| CorrDim | 78 | 83 | 84 | 82 |
| DANCo | 80 | 82 | 83 | 82 |
| Proposed method | 81 | 84 | 87 | 87 |

The proposed method was implemented after PCA while intrinsic dimension estimation methods were performed before PCA. This might be one of the reasons for the proposed method to perform better. By applying the proposed method after PCA, the proposed method evaluated the discriminative power of the transformed data in different number of dimensions. But the intrinsic estimation methods estimated the number of dimensions before knowing the discriminative power of the transformed data.

Besides, the performance of MLE, CorrDim and DANCO were highly depend on the k value. K nearest neighbor (KNN) was adopted in these techniques. A smaller k value may focus on the noisy data, and it is high impact towards the result. On the other hand, larger k value may cause bias and increase computation time. As a result, the proposed method outperformed the current intrinsic estimation methods.

### 4.2 The Comparison of Different Classifiers in Feature Extraction

The proposed technique used the combination of RSM and LDA to estimate the intrinsic dimension. However, KNN can be another choice for the weak learner. Therefore, this section includes the combination of RSM and KNN in comparison. Besides this, LDA and KNN also are compared in this section without RSM to investigate the strength of RSM. The k values for KNN were decided based on the cross validation on the classifier. The classification results by using different classifiers are shown in Table 4.

The average accuracy of KNN and LDA across the time points was same, which was 83%. The results in Table 4 convinces that RSM plays an important role in achieving a better result. RSM+KNN and RSM+LDA achieved higher results compared to use KNN or LDA alone, and RSM+LDA performed better compared to RSM+KNN. Besides, we examined the wrapper approach by using SVM to find the intrinsic dimension. Although SVM is a powerful machine learning approach, but the results showed that the proposed method outperformed it in feature extraction.

*Table 4: A Comparison of Proposed Classifier and Other Classifiers*

| Methods | 24m | 18m | 12m | Stable |
|---|---|---|---|---|
| KNN | 80 | 83 | 82 | 87 |
| RSM+KNN | 80 | 83 | 84 | 88 |
| LDA | 80 | 82 | 85 | 85 |
| SVM | 81 | 83 | 84 | 87 |
| Proposed method | 81 | 84 | 87 | 87 |

The reason for using LDA rather than KNN because LDA is a parameter free model. It will eliminate the concern on selecting parameter and reduce the computation time. Moreover, it is hardly saying which part of the brain provides more distinguishing power. Therefore, this study considers squeezing the most information together with PCA. After that, LDA became the classifier to decide the optimal number of features to distinguish the sick patient and healthy patient. Eventually, it forms the proposed technique, called PCRSDA.

PCRSDA uses cross validation to optimize the algorithm through RSM. By using RSM, the samples and the variables are selected randomly in every iteration, it reduces overdependence on the certain samples and variables. It allows the algorithm to evaluate the impact of numbers of dimensions

towards the classification results. The evaluation of each number of dimensions is done in 1000 times because the algorithm involves 10-fold cross validation and 100 learning cycles. Though the computation time of RSM+LDA is longer than only using LDA, but it is durable, and it is worth for the sake of increasing the accuracy.

## 5. CONCLUSION

A feature extraction method which consists of the combination of PCA with the specified design intrinsic dimension estimation has been proposed to solve the problem of extracting the information that well explained the data for Alzheimer's Diseases classification. PCA maximizes the variance of the data, which makes the most information to squeeze in first principal component. However, it is needed to include other principal components to give best explanation of the data towards Alzheimer's Disease. As the result, this study proposed RSM+LDA as the intrinsic dimension estimation approach for PCA. The numbers of dimensions were set from 1 to the dimension of the data to compare the classification error by using various dimensions. RSM selects the variables randomly for each LDA, which decreases the generalization error of using the fixed number of dimensions. By randomly choosing the variables, RSM also eliminates the function of filter approach. Therefore, there is no need to compare with the combination of filter approach with the proposed method. The findings show that RSM+LDA performed better than the current intrinsic dimension estimation. It extends our knowledge on the advantages of RSM in improving the results. Apart from that, cross validation also contributes to enhance the strength of RSM.

Despite the promising results, there are rooms to improve because the obtained accuracy still not the highest result of PCA. The strength of RSM is not fully discovered in this study. We suggest that RSM can be used as a tool to find the subset of principal components that gives better explanation of the progressive of AD. This is because the low variance principal components also can be the significant feature when it combined with other principal components. Besides, this study does not manage to benchmark with other papers out of two reasons. First, most of the paper did not provide the subject identification numbers of the dataset. Therefore, the researchers collected different samples even though the samples were from same database. Second, there are different features can be used in AD classification, and the classification result is highly dependent on the features. Therefore, this paper could not benchmark with the studies which used different features with this study. This study also only compared the proposed method with intrinsic estimation methods. Even though there are limitations on this study, but the validation results still proved that PCRSDA is a promising method for increasing the AD classification results.

## ACKNOWLEDGMENT:

The authors would like to thank Universiti Teknologi Malaysia (UTM) for supporting this research. The authors also acknowledge that the data used in this study was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) 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.

The data was funded by 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: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to 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 Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the

Laboratory for Neuro Imaging at the University of Southern California.

**REFRENCES:**

[1] B. Dubois, H. H Feldman, C. Jacova, S. T DeKosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha, K. Meguro, J. O'Brien, F. Pasquier, P. Robert, M. Rossor, S. Salloway, Y. Stern, P. J Visser, P. Scheltens," *Lancet Neurology*, vol. 6, no. 8. 2007, pp. 734–746.

[2] C. R. Munteanu, C. Fernandez-Lozano, V. Mato Abad, S. Pita Fernández, J. Álvarez-Linera, J. A. Hernández-Tamames, and A. Pazos, "Classification of mild cognitive impairment and Alzheimer's Disease with machine-learning techniques using 1H Magnetic Resonance Spectroscopy data," *Expert Syst. Appl.*, vol. 42, no. 15–16, 2015, pp. 6205–6214.

[3] F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," *Proc. 2016 IEEE Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2016*, 2016, pp. 2264–2268.

[4] C. Salvatore, A. Cerasa, and I. Castiglioni, "MRI Characterizes the Progressive Course of AD and Predicts Conversion to Alzheimer's Dementia 24 Months Before Probable Diagnosis," *Front. Aging Neurosci.*, vol. 10, May 2018.

[5] L. Khedher, J. Ramírez, J. M. Górriz, A. Brahim, and F. Segovia, "Early diagnosis of Alzheimer ' s disease based on partial least squares , principal component analysis and support vector machine using segmented MRI images," *Neurocomputing*, vol. 151, 2015, pp. 139–150.

[6] A. Demirhan, T. M. Nir, A. Zavaliangos-Petropulu, C. R. Jack, M. W. Weiner, M. A. Bernstein, P. M. Thompson, and N. Jahanshad, "Feature selection improves the accuracy of classifying Alzheimer disease using diffusion tensor images," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2015, vol. 20, no. 3, pp. 126–130.

[7] I. Beheshti and H. Demirel, "Feature-ranking-based Alzheimer's disease classification from structural MRI," *Magn. Reson. Imaging*, vol. 34, no. 3, 2016, pp. 252–263.

[8] E. Levina and P. J. Bickel, "Maximum Likelihood Estimation of Intrinsic Dimension," in *Proceedings of the 17th International Conference on Neural Information Processing Systems*, 2004, pp. 777–784.

[9] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Phys. D Nonlinear Phenom.*, vol. 9, no. 1–2, 1983, pp. 189–208.

[10] K. Fukunaga and D. R. Olsen, "An Algorithm for Finding Intrinsic Dimensionality of Data," *IEEE Trans. Comput.*, vol. C–20, no. 2, Feb. 1971, pp. 176–183.

[11] P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza, "Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework," *Math. Probl. Eng.*, vol. 2015, 2015, pp. 1–21.

[12] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, and P. Campadelli, "DANCo: An intrinsic dimensionality estimator exploiting angle and norm concentration," *Pattern Recognit.*, vol. 47, no. 8, 2014, pp. 2569–2581.

[13] J. Mielniczuk and P. Teisseyre, "Using random subspace method for prediction and variable importance assessment in linear regression," *Comput. Stat. Data Anal.*, vol. 71, 2014, pp. 725–742.

[14] R. Bryll, R. Gutierrez-Osuna, and F. Quek, "Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets," *Pattern Recognit.*, vol. 36, no. 6, 2003, pp. 1291–1302.

[15] T. J. Cleophas and A. H. Zwinderman, "Discriminant Analysis for Supervised Data," in *Machine Learning in Medicine*, Dordrecht: Springer Netherlands, 2013, pp. 215–224.

[16] K. Shen, J. Fripp, F. Mériaudeau, G. Chételat, O. Salvado, and P. Bourgeat, "Detecting global and local hippocampal shape changes in Alzheimer's disease using statistical shape models," *Neuroimage*, vol. 59, no. 3, 2012, pp. 2155–2166.

[17] H. Park, J. J. Yang, J. Seo, and J. M. Lee, "Dimensionality reduced cortical features and their use in the classification of Alzheimer's disease and mild cognitive impairment," *Neurosci. Lett.*, vol. 529, no. 2, 2012, pp. 123–127.

[18] A. Farzan, S. Mashohor, A. R. Ramli, and R. Mahmud, "Boosting diagnosis accuracy of Alzheimer's disease using High dimensional recognition of longitudinal brain atrophy patterns," *Behav. Brain Res.*, vol. 290, Apr. 2015, pp. 124–130.

[19] H. Park, J. ju Yang, J. Seo, and J. min Lee,

"Dimensionality reduced cortical features and their use in predicting longitudinal changes in Alzheimer's disease," *Neurosci. Lett.*, vol. 550, 2013, pp. 17–22.

[20] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach," *Int. J. Comput. Sci. Inf. Secur.*, vol. 12, no. 8, Sep. 2014, pp. 33–39.

[21] N. Bolse, A. Habermehl, C. Eschenbaum, and U. Lemmer, *Fluorescence quenching sensor arrays for the discrimination of nitroaromatic vapors*. 2018.

[22] T. K. Ho, "Random Decision Forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, pp. 278–282.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, Sep. 1995, pp. 273–297.

[24] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal, "Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity.," *Emergency*, vol. 3, no. 2, 2015, pp. 48–49.