Applying Classification on Structural MRI Images to Diagnose Alzheimer's Disease

Ali Hejazizo[1]

[1] *University of Alberta, Edmonton, Canada*

December 25, 2017

1 Introduction

Alzheimer's disease (AD) is a growing health problem and the most frequent neurodegenerative dementia. Early and accurate diagnosis of Alzheimer's Disease (AD) is not only challenging, but is crucial in the perspective of future treatments. Clinical diagnostic criteria are currently based on the clinical examination and neuropsychological assessment, with the identification of dementia and then of the Alzheimer's phenotype [3]. Studies suggest that volumetric measurements of regions of interests (ROI) in brain are useful to identify AD patients, separating them from normal elderly individuals [1].

As AD progresses, brain tissue shrinks and the ventricles, chambers within the brain that contain cerebrospinal fluid, become noticeably enlarged. In the final stages, people may lose the ability to feed themselves, speak, recognize people and control bodily functions. Memory

worsens and may become almost non-existent. On average, those with Alzheimer's live for 8 to 10 years after diagnosis, but this terminal disease can last for as long as 20 years.

1.1 Problem Statement

There has been considerable research toward the diagnosis and early detection of this disease in the past decade. Advances in statistical learning with the development of new machine learning algorithms that can handle high dimensional data, such as the support vector machine (SVM), helped to develop new diagnostic tools based on T1-weighted MRI to diagnose AD based on the volumetric measurement of different regions of the brain with high accuracy [2].

Using the volumetric measurements of regions of interest, the goal of this project is to:

1. Diagnose subjects with AD, based on MRI.
2. Compare different methods for the classification of patients, using the same study population.

For the studies performed in this project, the Open Access Series of Imaging Studies (OASIS) database is selected which is elaborated in section 2.

2 OASIS Dataset Exploration

The Open Access Series of Imaging Studies (OASIS) is a series of magnetic resonance imaging data sets that is publicly available for study and analysis. The initial data set consists of a cross-sectional collection of 416 subjects aged 18 to 96 years. One hundred of the included subjects older than 60 years have been clinically diagnosed with very mild to moderate Alzheimers disease. The subjects are all right-handed and include both men and women. For each subject, three or four individual T1-weighted magnetic resonance imaging scans are obtained in single imaging sessions. 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.

Dementia status is established using the clinical dementia rating (CDR) Scale. The CDR is a 5-point scale used to characterize six domains of cognitive and functional performance applicable to Alzheimer disease and related dementias. The necessary information to make each rating is obtained through a semi-structured interview of the patient and a reliable informant

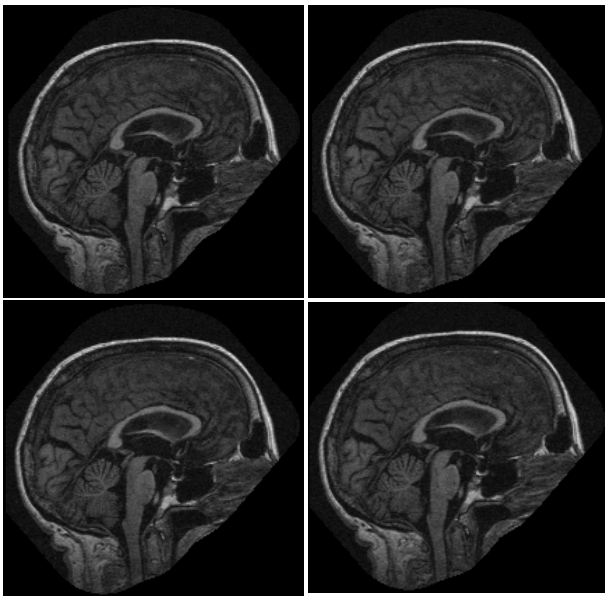

Figure (1) *individual T1-weighted magnetization prepared rapid gradient-echo (MP-RAGE) images of one random subject*

or collateral source (e.g., family member). In addition to ratings for each domain, an overall CDR score may be calculated through the use of an algorithm. To characterize and track a patient's level of impairment/dementia, this score is useful:

- 0 = Normal
- 0.5 = Very Mild Dementia
- 1 = Mild Dementia
- 2 = Moderate Dementia
- 3 = Severe Dementia

For each subject, 34 individual T1-weighted magnetization prepared rapid gradient-echo (MP-RAGE) images are acquired on a 1.5-T Vision scanner (Siemens, Erlangen, Germany) in a single imaging session. Which is shown in Figure 1 for one random subject.

Averaged motion-corrected images are then produced using the T1-weighted MP-RAGE images for each subject to improve signal-to-noise ratio, which is shown in Figure 2.

3 Data Preprocessing

To retrieve volumetric information of different parts of the brain, we have to perform several preprocessing stages, a few of which are mentioned below:

- Motion Correction and Conform

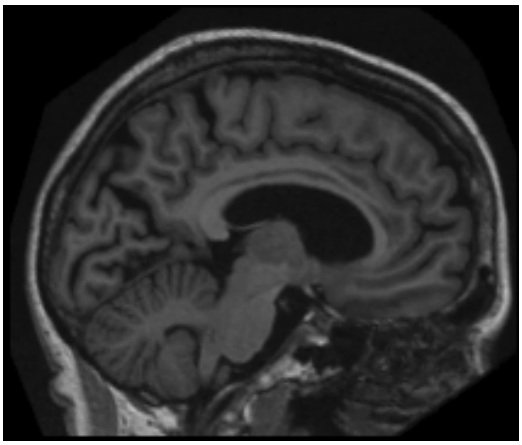

Figure (2) *Averaged motion-corrected image of a random subject acquired from images in Figure 1*

- NU (Non-Uniform intensity normalization)
- Talairach transform computation
- Intensity Normalization 1
- Skull Strip
- EM Register (linear volumetric registration)
- CA Intensity Normalization
- CA Non-linear Volumetric Registration
- Remove Neck
- LTA with Skull
- ...

In total 31 preprocessing stages is performed on each image with the FreeSurfer image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harva Details of each step is described in detail in free-surfer documentation. The result of preprocessing is shown in Figure 3 from sagittal, coronal, and transverse view. The preprocessed images are visualized using FSL tools which is a comprehensive library of analysis tools for FMRI, MRI and DTI brain imaging data. Figure 3 shows how perfectly the skull is stripped, neck is totally removed, etc.

Out of 416 subjects MRI images, 199 images successfully preprocessed. Other images excluded from the dataset due to poor quality of the original image or unknown CDR label.

3.1 Feature Extraction

Having preprocessed images, volume of different regions of the brain is extracted using the freesurfer tools. The features are obtained of brain segmentation and parcellation from both

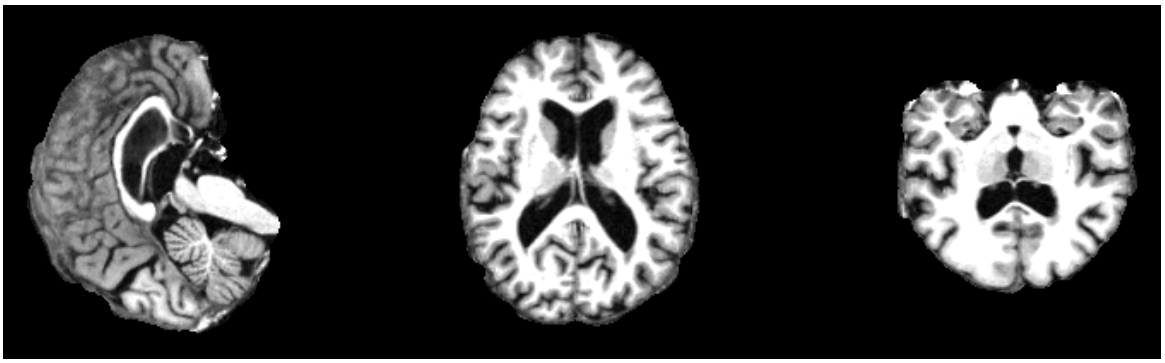

Figure (3) *Preprocessed image of the subject shown in Figure 2 in sagittal, transverse, and coronal views (left to right)*

Table (1) *OASIS Dataset*

Classes	2
Samples total	199
Samples per class	86 (1), 113(0)
Dimensionality	139
Features	Real values

left and right hemisphere. 139 features related to the volume of different regions of interest are extracted in total, some of which are mentioned below:

1. Left and right lateral ventricle
2. Left and right cerebellum white matter
3. Cerebrospinal fluid (CSF)
4. Left and right hippocampus
5. left and right hemisphere cortex
6. Estimated total intra cranial (eTIV)
7. left and right hemisphere surface holes
8. ...

Finally, as objective of the classification is to diagnose AD, the subjects with CDR greater than 0 are labeled as 1 (patient) and others (CDR = 0) are labeled as 0 (a control subject).

Table 1 shows the outline of the OASIS dataset.

4 Visualization

The dataset is almost a balanced dataset, i.e. the number of samples of patients and controlled subjects is balanced. Also it should be noted that the greatest known risk factor for Alzheimers is increasing age. Most individuals with the disease are 65 and older. Figure 4 shows the

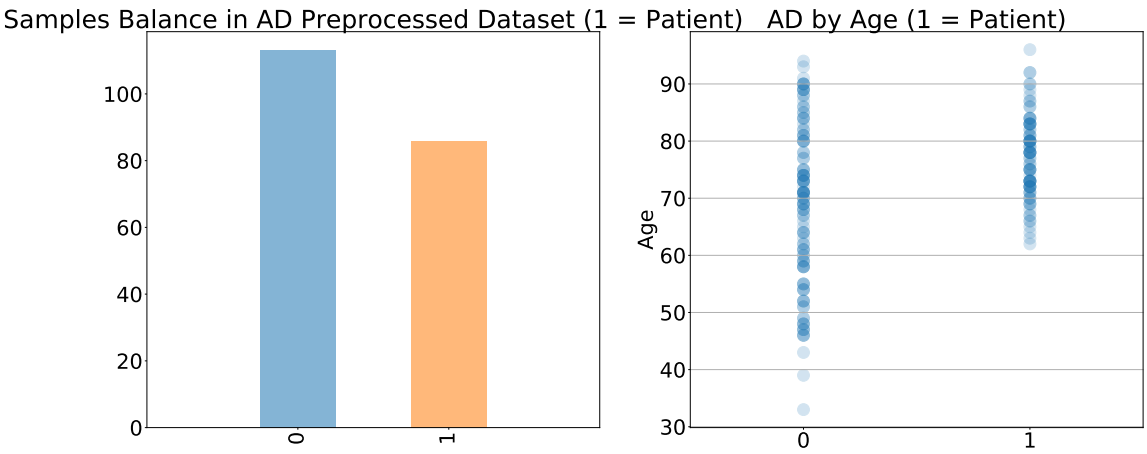

Figure (4) *Dataset balance and AD relation with age*

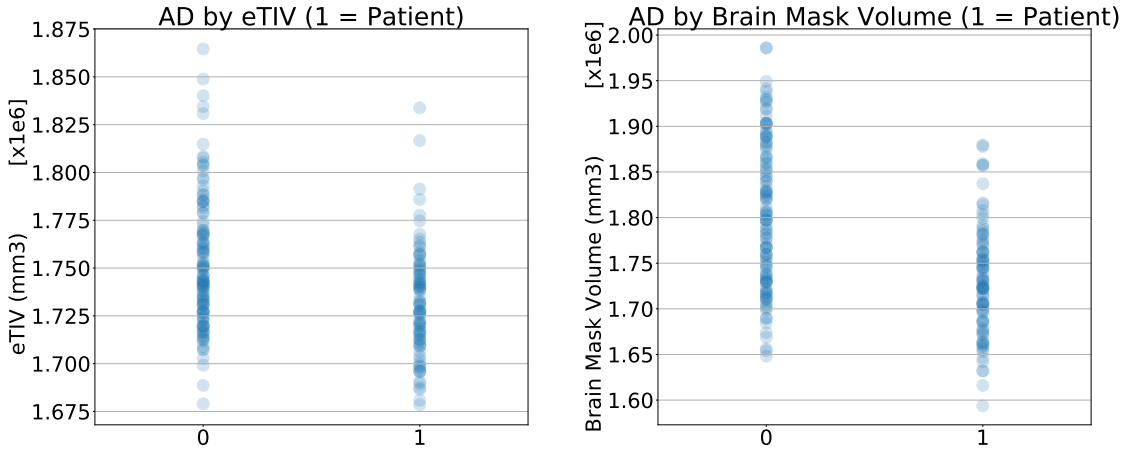

Figure (5) *eTIV and brain mask shrinkage in AD patients and contorlled subjects*

dataset balance on the left and the distribution of age in the dataset for patients and controlled subjects separately on the right.

As AD progresses, brain tissue shrinks. As an example, the eTIV and brain mask volume features extracted after preprocessing are shown in Figure 5. It can be seen from the figures that patients are likely to have smaller eTIV and brain mask which can be effectively used in identification of AD.

As another example, cerebral white matter and cortex volume are shown in Figure 6. The figure shown that the two features tend to have smaller value in patient compared to controlled subjects. If we use both features, it helps more to make patients separable from controlled subjects. This is shown if Figure 7. Compared to the individual cerebral white matter by age and cortex volume by age, patients and controlled subjects are more separable. Therefore, classification by using all the 139 volumetric features extracted after preprocessing helps to

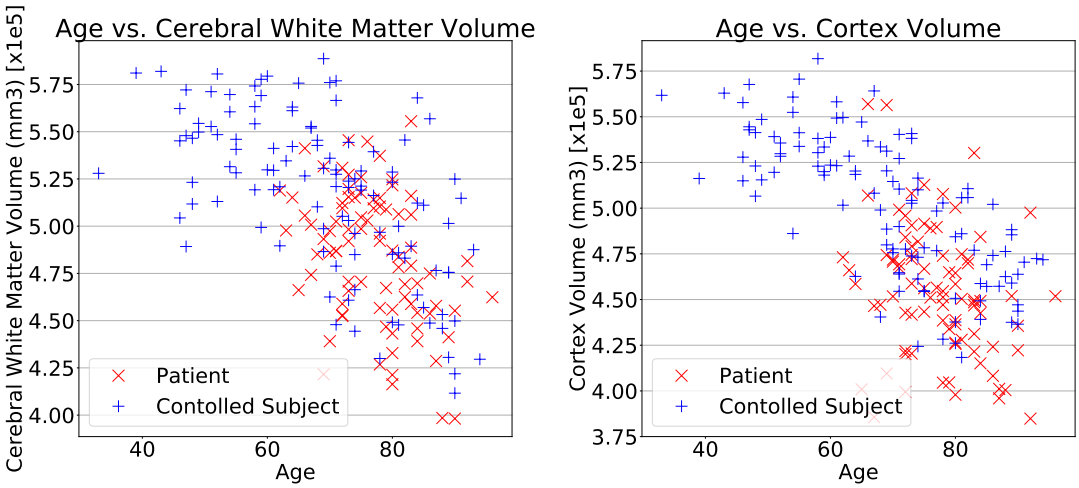

Figure (6) *AD relation with cerebral white matter volume, cortex volume, and age*

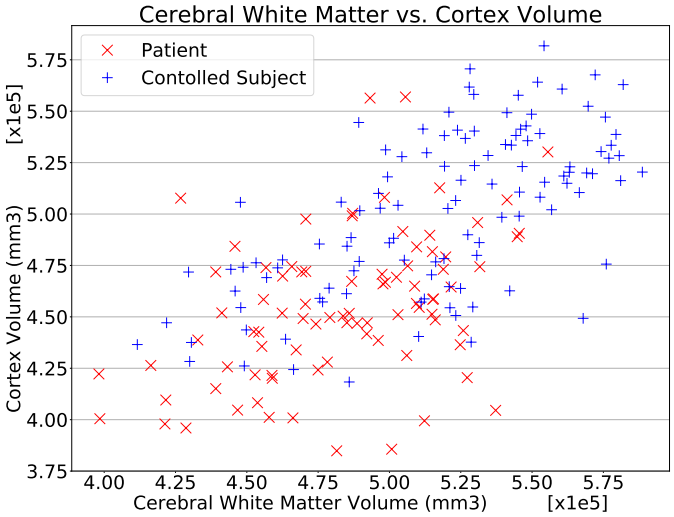

Figure (7) *Cerebral white matter volume vs. cortex volume in AD patients and controlled subjects*

achieve a high accuracy in AD identification.

In contrast, the ventricles, chambers within the brain that contain CSF, are noticeably enlarged in AD patients. This is shown in Figure 8 by plotting the CSF volume versus age.

5 Algorithms and Techniques

Data Analysis is done using the following supervised machine learning Techniques:

- Logistic Regression: Logistic regression deals with this problem by using a logarithmic transformation on the outcome variable which allow us to model a nonlinear association in a linear way. Therefore, logistic regression is chosen here to evaluate linear classifiers

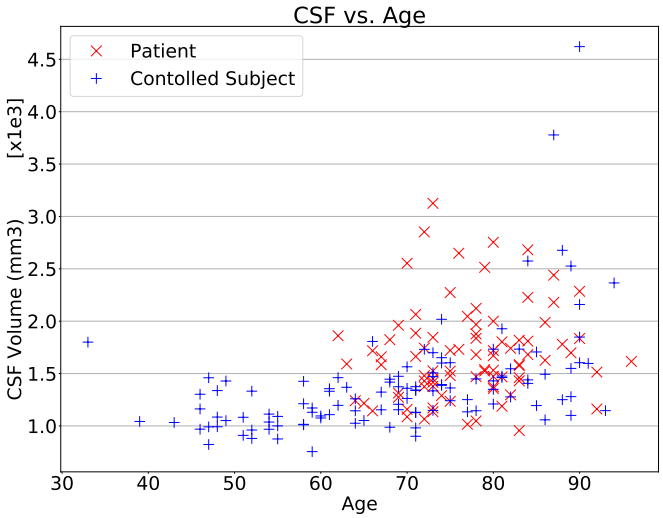

Figure (8) *CSF vs. age in AD patients and controlled subjects*

in OASIS dataset classification task.

- Random Forests: Random Forests perform implicit feature selection and provide a pretty good indicator of feature importance. As the number of features are relatively high, using random forest can help us to try considering only important features.
- SVM: The kernel trick in SVM is to transform data and then based on these transformations, find an optimal boundary between the possible outputs. Simply, it does some extremely complex data transformations, then figures out how to separate data based on the labels or outputs defined. The benefit is that we can capture complex relationships between data points without having to perform difficult transformations on our own. The downside is that the training time is much longer as it is much more computationally intensive.

5.1 Metrics

Classifiers are commonly evaluated using either a numeric metric, such as accuracy, or a graphical representation of performance, such as a receiver operating characteristic (ROC) curve.

A very simple choice to evaluate learning algorithms is the score which is the percentage of passengers correctly predicted. This is known simply as "accuracy. Other metrics such as area under the curve (AUC), recall, and precision are also considered which can be obtained from confusion matrix and ROC curve. Here recall is an important metric as it can be more detrimental to predict a patient is not sick if they are actually sick (False Negative), resulting

in a decision not to run further diagnostics and so causing serious complications from not treating the illness.

Two popular approaches for evaluating the performance of a classification algorithm on a data set are k-fold and leave-one-out cross validation. When the amount of data is large, k-fold cross validation should be employed to estimate the accuracy of the model induced from a classification algorithm, because the accuracy resulting from the training data of the model is generally too optimistic. Leave-one-out cross validation is a special case of k-fold cross validation, in which the number of folds equals the number of instances. When the number of instances either in a data set or for a class value is small, such as gene sequence data, leave-one-out cross validation should be adopted to obtain a reliable accuracy estimate for a classification algorithm. In this project as the number of instances is large enough, we study k-fold cross validation.

Then we apply statistical significant tests on the results obtained by k-fold cross validation. The following tests will be applied to compare different classifiers performance:

- Students t-test (the simplest statistical test)
- Paired tests

5.2 Implementation

The three classifiers are implemented in `python` using `sklearn` library.

6 Expriments and Results

The three algorithms mentioned in Section 5 are implemented to classify titanic dataset. The three algorithms are:

- Logistic regression
- Random Forests
- SVM classification

6.1 Hyperparameters tuning

For each classifier there are parameters that needs to be tuned in order to have a fair comparison. The hyperparameters that are tuned here is as follows:

- Logistic Regression
  - Regularization Parameter
    - `[1e-1, 1e-2, 1e-3, 1e-4, 1e-5]`
  - Tolerance
    - `[1e-1, 1e-2, 1e-3, 1e-4, 1e-5]`
- Random Forests
  - Number of estimators
    - `5, 10, 15, 20, 25`
  - Max depth
    - `[2-10]`
- SVM
  - C (l2 Regularization Coefficient)
    - `[0.1, 1, 10, 100, 1000]`
  - Gamma (Free parameter of the Gaussian radial basis function)
    - `[10, 1, 1e-1, 1e-2, 1e-3, 1e-4]`
  - Kernel type
    - `[Linear, Poly, rbf]`

7 Results

The algorithms were run over `199` samples with `139` features. Internal and external cross validation is applied with `K=10` for external and `k=5` for internal (parameter tunning cross validation).

The results for logistic regression, random forests, and SVM are shown in Tables 2, 3, and 4, respectively.

The metrics are as follows:

- Accuracy
- Confusion matrix
  - Recall
  - Precision

Table (2) *Logistic regression performance*

Logit	Train Accuracy	Test Accuracy	TN	FP	FN	TP	AUC	Recall	Precision
K_0	0.8202	0.7143	10	2	4	5	0.6944	0.5556	0.7143
K_1	0.809	0.8571	11	1	2	7	0.8472	0.7778	0.875
K_2	0.809	0.5238	8	4	6	3	0.5	0.3333	0.4286
K_3	0.7989	0.85	11	0	3	6	0.8333	0.6667	1
K_4	0.7933	0.8	9	2	2	7	0.798	0.7778	0.7778
K_5	0.8045	0.9	10	1	1	8	0.899	0.8889	0.8889
K_6	0.7667	0.7895	11	0	4	4	0.75	0.5	1
K_7	0.8222	0.6842	8	3	3	5	0.6761	0.625	0.625
K_8	0.8278	0.5789	6	5	3	5	0.5852	0.625	0.5
K_9	0.7833	0.8421	10	1	2	6	0.8295	0.75	0.8571

Table (3) *Random forests performance*

RandomFrest	Train Accuracy	Test Accuracy	TN	FP	FN	TP	AUC	Recall	Precision
K_0	0.8652	0.5714	9	3	6	3	0.5417	0.3333	0.5000
K_1	0.9101	0.8095	8	4	0	9	0.8333	1.0000	0.6923
K_2	0.8933	0.7143	10	2	4	5	0.6944	0.5556	0.7143
K_3	0.8994	0.7500	9	2	3	6	0.7424	0.6667	0.7500
K_4	0.8827	1.0000	11	0	0	9	1.0000	1.0000	1.0000
K_5	0.8436	0.7000	8	3	3	6	0.6970	0.6667	0.6667
K_6	0.8722	0.7368	8	3	2	6	0.7386	0.7500	0.6667
K_7	0.8722	0.7368	7	4	1	7	0.7557	0.8750	0.6364
K_8	0.8833	0.7895	9	2	2	6	0.7841	0.7500	0.7500
K_9	0.8778	0.8421	10	1	2	6	0.8295	0.7500	0.8571

Table (4) *Support Vector Machine performance*

SVM	Train Accuracy	Test Accuracy	TN	FP	FN	TP	AUC	Recall	Precision
K_0	0.8539	0.619	8	4	4	5	0.6111	0.5556	0.5556
K_1	0.8652	0.9048	11	1	1	8	0.9028	0.8889	0.8889
K_2	0.8764	0.6667	6	6	1	8	0.6944	0.8889	0.5714
K_3	0.8771	0.85	11	0	3	6	0.8333	0.6667	1
K_4	0.8603	0.7	7	4	2	7	0.7071	0.7778	0.6364
K_5	0.8547	0.65	7	4	3	6	0.6515	0.6667	0.6
K_6	0.8889	0.7368	10	1	4	4	0.7045	0.5	0.8
K_7	0.8389	0.9474	10	1	0	8	0.9545	1	0.8889
K_8	0.85	0.7368	9	2	3	5	0.7216	0.625	0.7143
K_9	0.8667	0.7368	9	2	3	5	0.7216	0.625	0.7143

Table (5) *Standard deviation on different criterions*

std	Logit	RF	SVM
Test Accuracy	0.125786260069	0.110474644653	0.110463227164
AUC	0.126781290681	0.117883898156	0.110411857254
Precision	0.19778777094	0.133647423802	0.152583119789
Recall	0.160892658834	0.200997872017	0.162600903373

- AUC (Area Under ROC Curve)

After the metrics were obtained across the 10 outer folds, in order to determine whether there is a significant difference between the mean of the three algorithms, an Analysis of variance (ANOVA) test was performed for each of the metrics (accuracy, recall, precision, AUC). The null and alternate hypotheses are as follows:

- **Null hypothesis**: The true mean of all the three algorithms are the same.
- **Alternate hypothesis**: The true mean of the three algorithms are different.

7.1 ANOVA test

ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups. ANOVAs are useful for testing three or more means (groups or variables) for statistical significance.

The ANOVA test has an important assumption that must be satisfied in order for the associated p-value to be valid:

- The population standard deviations of the groups are all equal. This property is known as homoscedasticity.

For that reason, we have to first calculate the standard deviation of the three groups of accuracy. The results are shown in Table

As the standard deviation values are almost equal, we can perform ANOVA test. The results of ANOVA test is shown in Table

The ANOVA test results demonstrate that the three classifiers are not significantly different and the null hypothesis can not be rejected. If the ANOVA test on any of the metrics was almost less than 0.1, we could go further and perform t-test to detect statistically different pairs of classifiers. Nevertheless, we have performed Welch t-test in the next section to further prove that the classifiers are not significantly different.

Table (6) *ANOVA test on different criterions*

Test/Criterion	ANOVA
Test Accuracy	0.9719575709
AUC	0.967891956479
Precision	0.969041006168
Recall	0.978508571491

Table (7) *Welch t-test*

Welch	Test Accuracy	AUC	Precision	Recall
Logit - SVM	0.987514558976	0.867896239608	0.711422776649	0.349726225454
RF - SVM	0.838589915295	0.825444904011	0.83410950904	0.85392449793
RF - Logit	0.837006408275	0.71377011516	0.573150893663	0.311849231929

7.2 T-Test

We calculate the T-test for the means of the pairs of two independent samples of scores, i.e. the scores obtained for algorithms.

The test measures whether the average (expected) value differs significantly across samples. If we observe a large p-value, for example larger than 0.05 or 0.1, then we cannot reject the null hypothesis of identical average scores. If the p-value is smaller than the threshold, e.g. 1%, 5% or 10%, then we reject the null hypothesis of equal averages.

As Welch's t-test performs better than Student's t-test whenever sample sizes and variances are unequal between groups, and gives the same result when sample sizes and variances are equal, we use Welch's t-test. We also provide the results of and standard t-test. The results for both tests, Welch and standard t-test, are shown in Tables 7 and 8, respectively. The results show the fact that when the variances are almost equal, Welch's t-test and standard t-test give similar results.

Considering the results of Welch's t-test, Table 7, we demonstrate that the null hypothesis cannot be rejected as the `p-value` is greater than 0.1 for any pair of algorithms and any metrics.

Table (8) *Standard t-test*

Standard	Test Accuracy	AUC	Precision	Recall
Logit - SVM	0.987517448177	0.867931453192	0.711703149617	0.349727624359
RF - SVM	0.83858991532	0.825455591706	0.834150603644	0.854014868845
RF - Logit	0.837045132973	0.713792872878	0.574128015242	0.312499517608

8 Conclusion

The results confirm that the volumetric measurements of different regions of the brain can be effectively used in AD identification. The experiments result also demonstrates that the original hypothesis that one of the algorithm chosen would not perform better than the other cannot be rejected through the experiments that have been done. In terms of the metrics chosen for comparison, the three algorithms performed similar, however it should be noted that the run time of SVM is significantly higher than the other two which is a benefit for logistic regression and random forests. The data preprocessing part also plays an important role in the quality of final results and accuracy. It may also affect the final results of classifier differently and therefore should be taken into account. As a result, Further work on preprocessing data can be considered as an important step in future work.

For future work, collecting more data of both controlled subjects and patients can help to achieve higher accuracy. Combining OASIS with Alzheimers Disease Neuroimaging Initiative (ADNI) dataset, which is another famous dataset in Alzheimer's disease, might help. Nevertheless, we should be aware of different scanning procedures and scaling as well as applying the same preprocessing and feature extraction on both dataset. Applying outliers detection can also be helpful, thereby detecting wrong labeled data to remove them from training set and finally, diagnosis accuracy improvement.

References

[1] Cássio MC Bottino, Cláudio C Castro, Regina LE Gomes, Carlos A Buchpiguel, Renato L Marchetti, and Mário R Louzã Neto. Volumetric mri measurements can differentiate alzheimer's disease, mild cognitive impairment, and normal aging. *International Psychogeriatrics*, 14(1):59–72, 2002.

[2] Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, Olivier Colliot, Alzheimer's Disease Neuroimaging Initiative, et al. Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. *neuroimage*, 56(2):766–781, 2011.

[3] Lidia Glodzik, Lisa Mosconi, Wai Tsui, Susan de Santi, Raymond Zinkowski, Elizabeth Pirraglia, Kenneth E Rich, Pauline McHugh, Yi Li, Schantel Williams, et al. Alzheimer's disease markers, hypertension, and gray matter damage in normal elderly. *Neurobiology of aging*, 33(7):1215–1227, 2012.