Alzheimer's Disease Diagnosis based on Structural MRI with Machine Learning Techniques

Ali Hejazizo, hejazizo@ualberta.ca

Abstract—There is not a specific test to diagnose Alzheimer's disease (AD). Its diagnosis should be based upon clinical history, neuropsychological and laboratory tests, neuroimaging and electroencephalography (EEG). Therefore, new approaches are necessary to enable earlier and more accurate diagnosis and to follow treatment results

In this study we used Machine Learning techniques for AD diagnosis. We studied 199 subject of which 86 subjects were diagnosed to have AD. Three-dimensional T1-weighted magnetic resonance images of each subject were parcellated into regions of interest (ROIs). Based upon the volumetric characteristics extracted from each ROI, we use three different classifiers to classify the subjects and finally evaluate them in classification of whole-brain anatomical magnetic resonance imaging to discriminate between patients with AD and control subjects. The results demonstrate the effectiveness of using volumetric measurements to diagnose AD with high accuracy which provides a potential for early diagnosis of Alzheimers disease.

Index Terms—Alzheimer's disease, magnetic resonance imaging, machine learning, classification

I. Introduction

Dementia is a growing health problem, and Alzheimer's disease (AD) is the most frequent neurodegenerative dementia in the elderly accounting for 50–60% of all cases [?], [?], [?]. AD patients benefit from early cholinesterase inhibitors [?], [?] and would consequently gain from early and accurate diagnosis of AD. In recent years, the early clinical signs of AD have been extensively investigated, leading to the concept of amnestic mild cognitive impairment (MCI) [?], [?]. However, early and accurate diagnosis of Alzheimer's Disease (AD) is not only challenging, but is crucial in the perspective of future treatments. clinical examination and neuropsychological assessment are two clinical diagnostic criteria with the identification of dementia and then of the Alzheimer's phenotype [?].

Besides neuropsychological examination, structural imaging is increasingly used to support the diagnosis of AD. Studies have shown that neurodegeneration in AD begins in the medial temporal lobe, successively affecting the entorhinal cortex, the hippocampus, the limbic system, then extending toward neocortical areas [?]. Therefore, the detection of medial temporal lobe atrophy (MTA) has received considerable effort and attention, and particularly in the hippocampus, the entorhinal cortex, and the amygdala [?], [?]. Visual rating scales, linear or volumetric measurements, and voxel-based approaches has been used to evaluate MTA. Overall, the sensitivity and specificity of hippocampus measurements for distinguishing AD patients from healthy aged subjects have been evaluated to range from 80% to 95% [?], [?], [?], [?], [?]. However, evaluation of MTA for diagnostic purposes in AD has limitations. MTA measurements are much less efficient in the pre-dementia conditions such as amnestic MCI [?], [?], [?], [?]. Atrophy in early stages of AD is not confined to the hippocampus or the entorhinal cortex. Other areas are affected in AD patients and MCI patients as well [?]. Whole-brain methods for characterizing brain atrophy may therefore be more efficient in differentiating AD and MCI patients who will evolve toward AD from healthy control subjects.

Studies suggest that volumetric measurements of regions in brain are useful to identify AD patients, separating them from normal elderly individuals [?]. 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.

A. 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 [?].

In this project, our purpose is to:

1) Individually classify AD patients and healthy elderly control subjects by using a whole-brain MR image analysis.
2) Compare different classifiers in AD diagnosis, using the same study population.

For the classification purpose, we parcellate the subjects' brain MRI into regions of interest (ROIs) and focus on characteristics of the distribution of the gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), which intuitively makes sense when dealing with neurodegenerative diseases in general and AD in particular. The parcellation result is used to train the machine learning classifiers and then perform prediction on unseen data.

II. Dataset Exploration

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

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 Alzheimer's 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 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, 3–4 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.

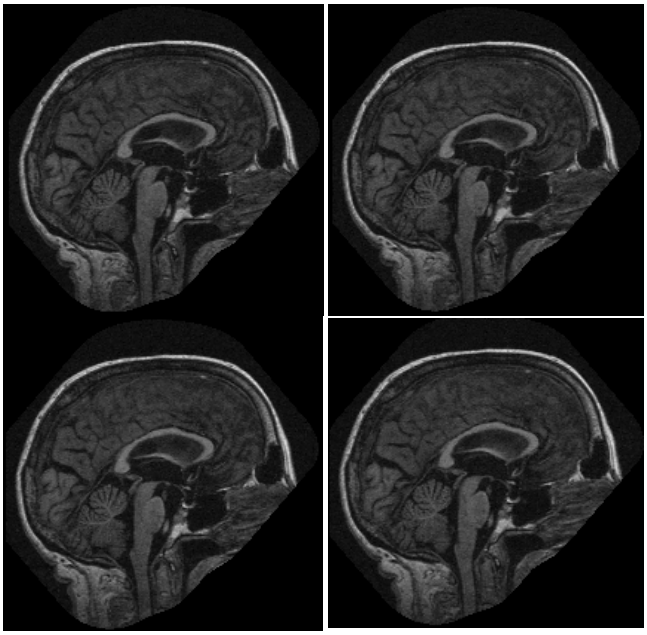

Fig. 1: individual T1-weighted magnetization prepared rapid gradient-echo (MP-RAGE) images of one random subject

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.

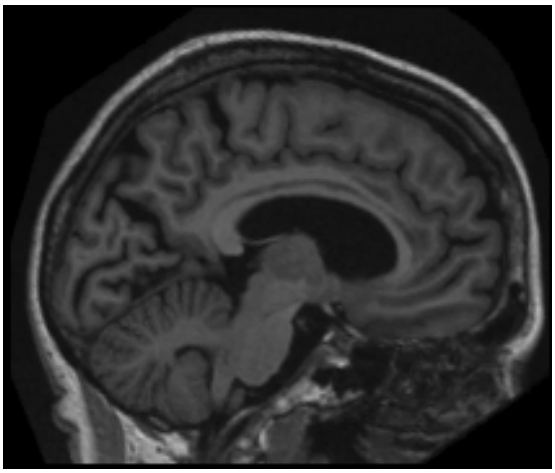

Fig. 2: Averaged motion-corrected image of a random subject acquired from images in Figure 1

III. 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
- 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.harvard.edu/). Details of each step is described 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 failed due to due to the poor

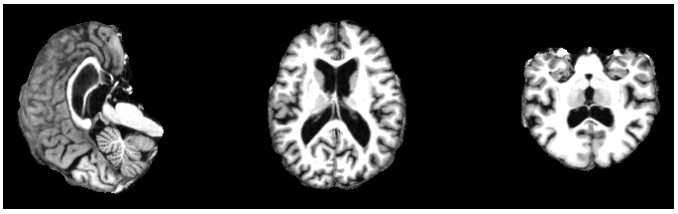

Fig. 3: Preprocessed image of the subject shown in Figure 2 in sagittal, transverse, and coronal views (left to right)

TABLE I: OASIS Dataset

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

quality of the original image or unknown CDR label.

A. Feature Extraction

Having preprocessed images, the brain is parcellated into ROIs and then volume of different regions of the brain is extracted using the freesurfer tools. The features are obtained of brain segmentation and parcellation from both 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 I shows the outline of the OASIS dataset.

IV. Visualization

The dataset is almost a balanced dataset, i.e. the number of samples of patients and controlled subjects is balanced. Figure 4 shows the dataset balance.

The greatest known risk factor for Alzheimer's is increasing age. Most individuals with the disease are 65 and older. This is shown in Figure 5 as the distribution of age in the dataset for patients and controlled subjects.

As AD progresses, brain tissue shrinks. As an example, the brain mask volume and eTIV features extracted after preprocessing are shown in Figures 6 and 7, respectively. 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.

Cerebral white matter and cortex volume are plotted in Figure 8. The figure shows that the two features tend

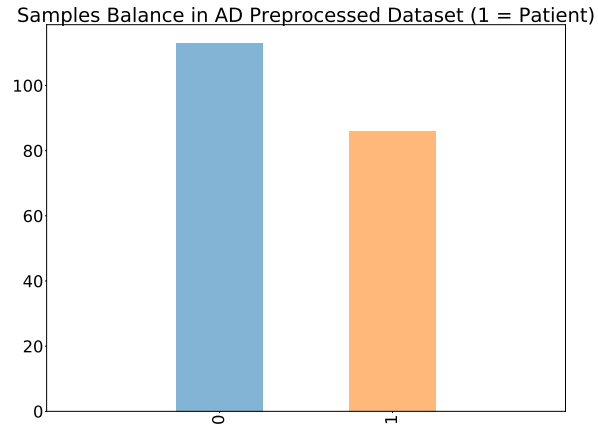

Fig. 4: Dataset balance

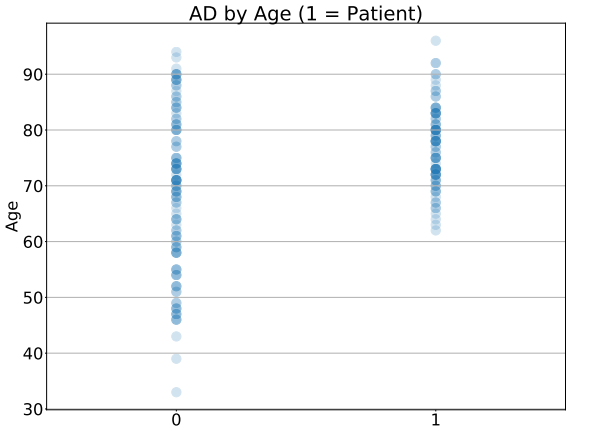

Fig. 5: Distribution of AD with age

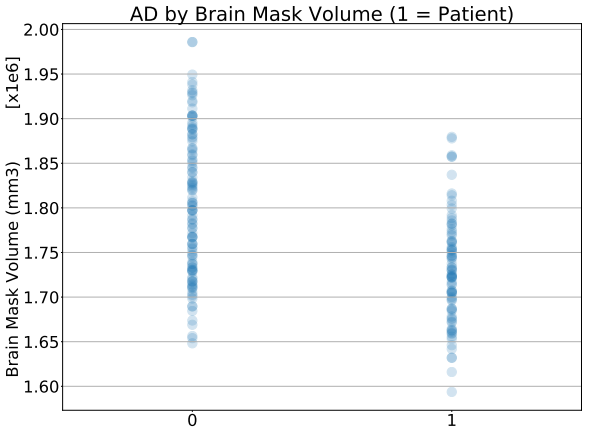

Fig. 6: Brain mask shrinkage in AD patients and contorlled subjects

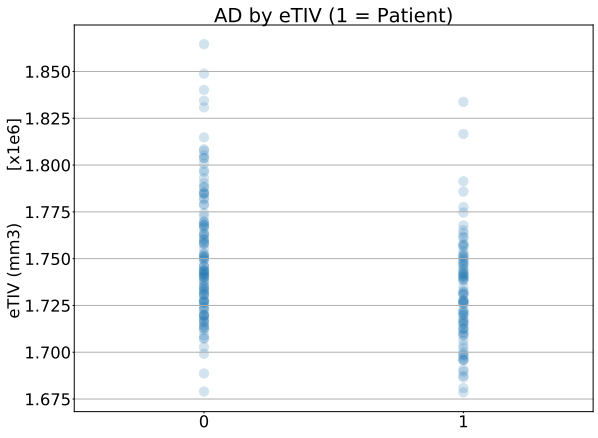

Fig. 7: eTIV shrinkage in AD patients and contorlled subjects

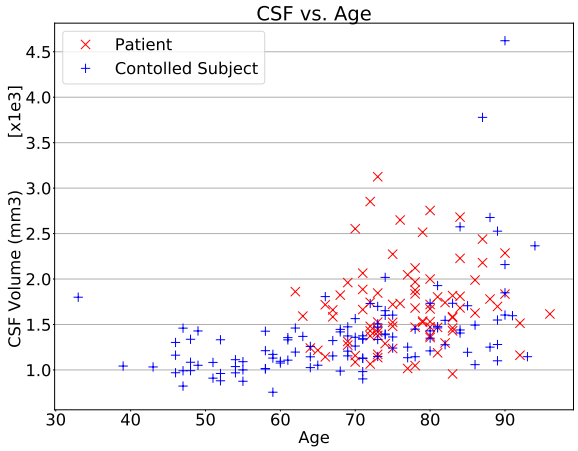

Fig. 9: CSF vs. age in AD patients and controlled subjects

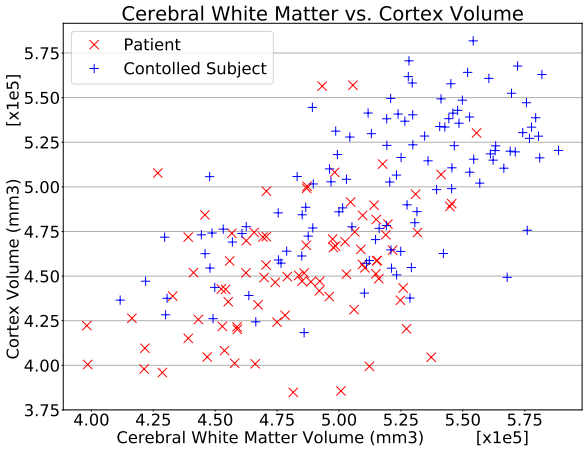

Fig. 8: Cerebral white matter volume vs. cortex volume in AD patients and controlled subjects

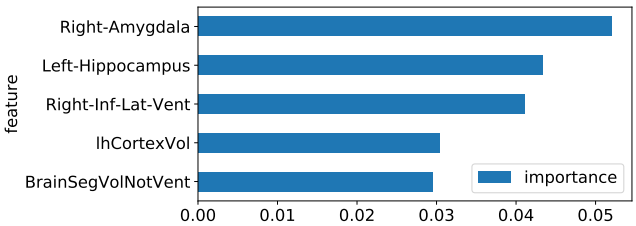

Fig. 10: Five most important features

to have smaller value in patient compared to controlled subjects. It can be seen that Patients and controlled subjects are relatively separable.

In contrast, the ventricles, chambers within the brain that contain CSF, are noticeably enlarged in AD patients. This is shown in Figure 9 by plotting the CSF volume versus age.

Plotting more important features in terms of how important they are in separating patients from control subjects, helps to achieve a better visualization. Therefore, we first calculate the features importance using random forest. The result is shown for the five most important features in Figure 10.

A 3D visualization of three most important features namely Right-Amygdala, Left-Hippocampus, and Right-Inf-Lat-Vent is shown in Figure 11. It can be clearly seen that the patient and controlled subjects have become more separable compared to the Figures 8 and 9.

Therefore, classification by using all the 139 volumetric features extracted after preprocessing helps to achieve a

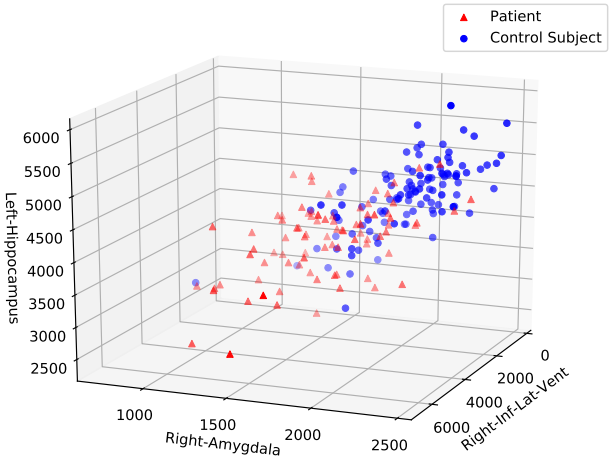

Fig. 11: Right-Amygdala, Left-Hippocampus, and Right-Inf-Lat-Vent 3D visualization

high accuracy in AD identification.

V. 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 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.

A. 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:

- Student's t-test (the simplest statistical test)
- Paired tests

B. Implementation

The three classifiers are implemented in python using sklearn library.

VI. Experiments

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

- Logistic regression
- Random Forests
- SVM classification

A. 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]

VII. 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 II, III, and IV, respectively.

The metrics are as follows:

- Accuracy

TABLE II: 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 III: 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 IV: 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

- Confusion matrix
  - Recall
  - Precision
- 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.

A. 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.

TABLE V: Standard deviation on different criterion

std	Logit	RF	SVM
Test Accuracy	0.1258	0.1105	0.1105
AUC	0.1268	0.1179	0.1104
Precision	0.1978	0.1336	0.1526
Recall	0.1609	0.2010	0.1626

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

TABLE VI: ANOVA test on different criterion

Test/Criterion	ANOVA
Test Accuracy	0.9720
AUC	0.9679
Precision	0.9690
Recall	0.9785

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.

B. 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 VII and VIII, 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 VII, 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.

VIII. Conclusion

Different classifiers that are able to classify patients with early Alzheimer's disease from control subjects based on the volumetric measurements of MRI images are developed and evaluated. Our results indicate that machine learning techniques can aid the clinical diagnosis of AD. The results confirm that the volumetric measurements of different regions of the brain can be effectively used in AD identification and provide a potential for early diagnosis of Alzheimer's disease.

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 Alzheimer's 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. It is also very useful to consider the CDR label as the output class and evaluate the methods in correctly differentiating between different forms of dementia which is more challenging that only AD diagnosis.

TABLE VII: 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

TABLE VIII: 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