FINDING AN OPTIMAL DIAGNOSTIC RECIPE FOR DETECTING THE AD PHASE

Problem Statement

Alzheimer's disease (AD) is a form of dementia that causes problems with memory, cognitive executive function, and behavior. AD is an irreversible process and is currently incurable. In order to make an accurate, early diagnosis, biomedical scientists have developed a complex diagnostic protocol which includes hundreds of tests that many affected/elderly participants are not able to finish. We want to address this issue by finding an optimal subset of the battery of tests that can perform as well, and be able to accurately detect the stage of the AD in the patient.

Project Goal

ADNI dataset includes 8 types of diagnoses (shown in Fig. 1). Each type contains multiple tests which are scored based on an AD Clinical Standard. From the diagnosis protocol, we find that some lengthy and difficult tests make the patients uncomfortable. There is a huge portion of AD patients who have abandoned their routinely diagnoses because of the unpleasant experience in the tests. From our data, we saw that 81 out of 3874 patients quit after being enrolled in the diagnosis program for 6 months. Among the 3793 patients left, only 730 had completed all required tests. By the end of the fourth year, the program was left with only 101 participants, which was originally planned to span over 5 years.

Motivated by the current state of affairs of the trials, our focus for this machine learning project was to study the correlation between the diagnosis of patients and their phase of AD. In the project, we took the test result of the patients as the attributes and the phase of AD (normal, mild, moderate, severe) as the classes. We aim to:

1. Implement machine learning models on the dataset to select the attributes (diagnostic tests) which are the most helpful in predicting the classes (phase of AD), thereby simplifying the current diagnostic recipe.
2. Find out how important attributes change over time, and suggest practical AD diagnostic recipes for different phases.

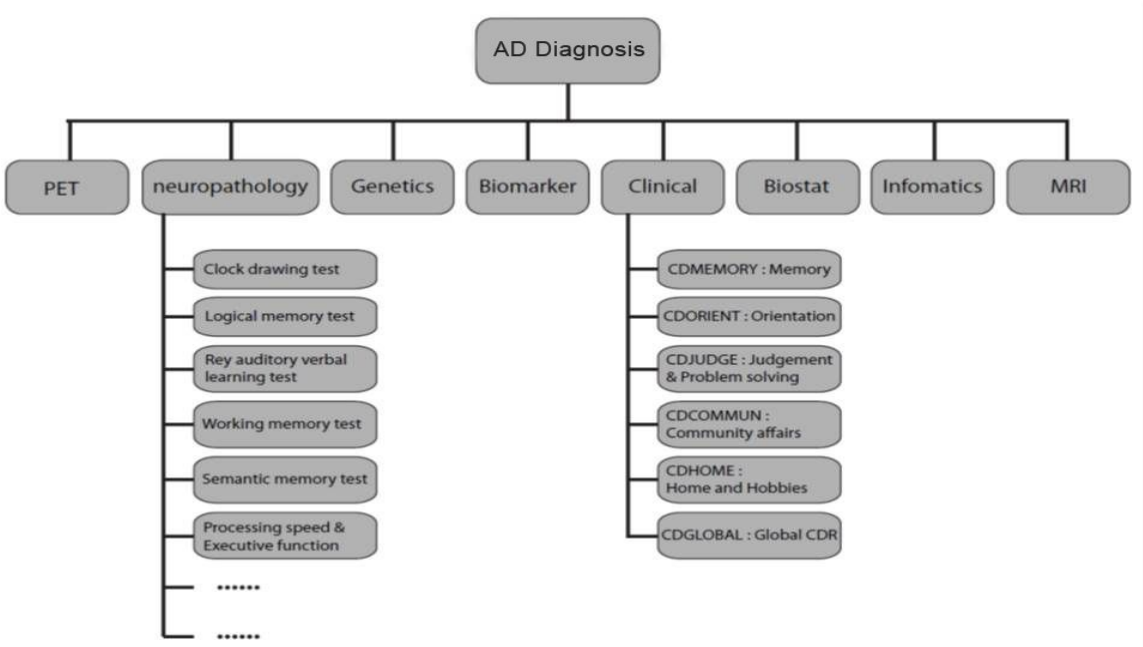


Figure 1 AD diagnosis in clinical medicine

Models and Analyses

Step 1: Cleaning the data

The *Alzheimer's Disease Neuroimaging Initiative* (ADNI) database has collected the diagnosis results from thousands of AD patients over years. It provides a good data resource to study AD. The evaluation of AD is based on the scores of the independent diagnoses. Each diagnosis includes tens of subcategories and an aggregated score for the category is generated by summarizing the scores of each subcategory. Currently, Clinical Dementia Rating (CDR) and Mild Cognitive Impairment (MCI) scores are used to evaluate the disease, while Mini-Mental State Examination (MMSE) score is used to define the phase of AD.

The raw data contains a large amount of missing data, which appear for various reasons. For example, the patients may not have been on-site for a scheduled test or chose to abandon the diagnosis program. We first excluded the missing data, and then collated different test entries using the patient ID (a unique identifier for each participant) using MATLAB. Our training data covers the diagnosis of 2752 patients over 4 years (shown in Table 1, M = month) and includes 4 classifiers and more than 50 attributes (shown in Table 2).

	M06	M12	M18	M24	M36	M48
No. Patients	730	675	305	554	387	101
No. Attributes	63	63	63	63	63	63

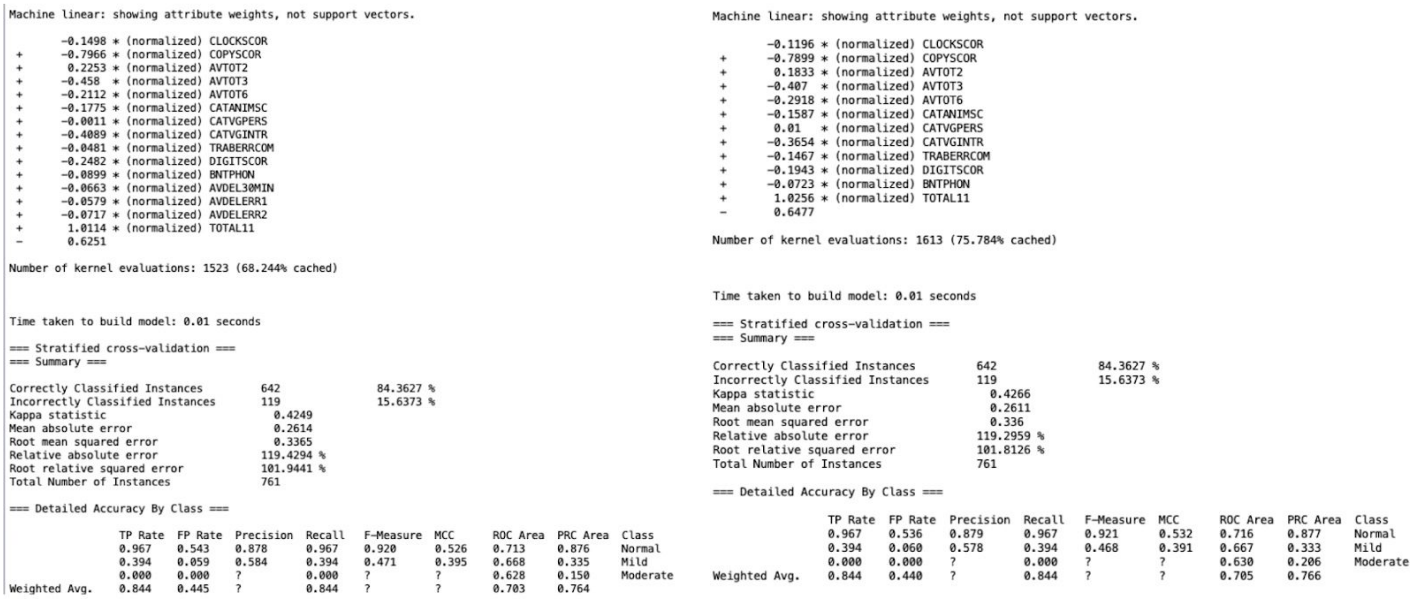
Table 1 Training data overview

Classifier	Normal, Mild, Moderate, Severe	
Attributes	CDMEMORY	Score of memory test
	CDORIENT	Score of orientation test
	CDJUDGE	Score of judgement & problem solving
	CDCOMMUN	Score of community affairs)
	CDHOME	Score of home hobbies
	CDCARE	Score of personal care
	CDGLOBAL	Score of global Clinical Dementia Rating)
	CLOCKCIRC, CLOCKSYM, CLOCKNUM, CLOCKHAND, CLOCKTIME, CLOCKSCOR, CDORIENT	Scores of clocking drawing test
	LMSTORY, LIMMTOTAL, LIMMEND	Scores of Logical memory test
	AVTOT1, AVTOT2, AVTOT3, AVTOT4, AVTOT5, AVTOT6, AVTOTB, AVERR1, AVERR2, AVERR3, AVERR4, AVERR5, AVERR6, AVERRB, AVENDED	Scores of Rey auditory verbal learning test (episodic memory test)
	DSPANFOR, DSPANFLTH, DSPANBAC, DSPANBLTH	Scores of working memory test, in which the subject is read number sequences of increasing length and asked to repeat them.
	CATANIMSC, CATANPERS, CATANINTR, CATVEGESC, CATVGPERS, CATVGINTR	Scores of semantic memory test (verbal fluency, language)
	TRAASCOR, TRAAERRCOM, TRAAERROM, TRABSCOR, TRABERRCOM, TRABERROM	Scores of Test of processing speed and executive function
	BNTND, BNTSPONT, BNTSTIM, BNTCSTIM, BNTPHON, BNTCPHON, BNTTOTAL	Scores of Boston naming test – a measure of the ability to orally label 30 line drawing of objects

Table 2 Classifiers and features in training model

Step 2: Selection of important attributes

We use the Decision Tree (J48) algorithm in Weka on the dataset of each month to select the attributes with the maximum information gain. If there are less than 12 leaves in the pruned tree, we keep all the attributes, otherwise choose the top 12 nodes. We have also tried different numbers of attributes to check how the accuracy changes, but it didn't show much improvement beyond 12 (Figure 1).



```
Machine linear: showing attribute weights, not support vectors.

       -0.1498 * (normalized) CLOCKSCOR
 +     -0.7966 * (normalized) COPYSCOR
 +      0.2253 * (normalized) AVTOT2
 +     -0.458  * (normalized) AVTOT3
 +     -0.2112 * (normalized) AVTOT6
 +     -0.1775 * (normalized) CATANIMSC
 +     -0.0011 * (normalized) CATVGPERS
 +     -0.4089 * (normalized) CATVGINTR
 +     -0.0481 * (normalized) TRABERRCOM
 +     -0.2482 * (normalized) DIGITSCOR
 +     -0.0899 * (normalized) BNTPHON
 +     -0.0663 * (normalized) AVDEL30MIN
 +     -0.0579 * (normalized) AVDELERR1
 +     -0.0717 * (normalized) AVDELERR2
 +      1.0114 * (normalized) TOTAL11
 -      0.6251

Number of kernel evaluations: 1523 (68.244% cached)



Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         642               84.3627 %
Incorrectly Classified Instances       119               15.6373 %
Kappa statistic                          0.4249
Mean absolute error                      0.2614
Root mean squared error                  0.3365
Relative absolute error                119.4294 %
Root relative squared error            101.9441 %
Total Number of Instances              761

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.967    0.543    0.878      0.967    0.920      0.526    0.713     0.876     Normal
               0.394    0.059    0.584      0.394    0.471      0.395    0.668     0.335     Mild
               0.000    0.000    ?          0.000    ?          ?        0.628     0.150     Moderate
Weighted Avg.  0.844    0.445    ?          0.844    ?          ?        0.703     0.764
```

```
Machine linear: showing attribute weights, not support vectors.

       -0.1196 * (normalized) CLOCKSCOR
 +     -0.7899 * (normalized) COPYSCOR
 +      0.1833 * (normalized) AVTOT2
 +     -0.407  * (normalized) AVTOT3
 +     -0.2918 * (normalized) AVTOT6
 +     -0.1587 * (normalized) CATANIMSC
 +      0.01   * (normalized) CATVGPERS
 +     -0.3654 * (normalized) CATVGINTR
 +     -0.1467 * (normalized) TRABERRCOM
 +     -0.1943 * (normalized) DIGITSCOR
 +     -0.0723 * (normalized) BNTPHON
 +      1.0256 * (normalized) TOTAL11
 -      0.6477

Number of kernel evaluations: 1613 (75.784% cached)



Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         642               84.3627 %
Incorrectly Classified Instances       119               15.6373 %
Kappa statistic                          0.4266
Mean absolute error                      0.2611
Root mean squared error                  0.336
Relative absolute error                119.2959 %
Root relative squared error            101.8126 %
Total Number of Instances              761

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.967    0.536    0.879      0.967    0.921      0.532    0.716     0.877     Normal
               0.394    0.060    0.578      0.394    0.468      0.391    0.667     0.333     Mild
               0.000    0.000    ?          0.000    ?          ?        0.630     0.206     Moderate
Weighted Avg.  0.844    0.440    ?          0.844    ?          ?        0.705     0.766
```

Figure 2. *Comparison of accuracies of SMO with 12 attributes and 15 attributes in Weka using 10-fold CV*

We found that as time passes, the number of important attributes (nodes in the tree) decreases. This suggests that as the symptoms of AD worsen, many of the more common tests may not be strong indicators of the state of the disease. For example, the test "AVDEL30MIN" (shown in gray), a difficult test that takes over 30 minutes, only exists in the top 12 in M06, while the test "CDORIENT" (shown in blue) appears for M12, M18, M24 and M36, indicating it is one of the strongest indicators of the severity of the condition.

	M06	M12	M18	M24	M36	M48
Selected tests	CDGLOBAL	TOTALMOD	TOTALMOD	CDORIENT	CDORIENT	DIGITSCOR
	BNTSPONT	CATVGINTR	DSPANBLTH	CLOCKHAND	TOTALMOD	TOTAL11
	AVDEL30MIN	CDSOURCE	AVERR3	DSPANBAC	TRABERRCOM	AVDELTOT
	CLOCKNUM	CLOCKSYM	TRAAERRCOM	TOTAL11	CLOCKNUM	CATANPERS
	TOTAL11	CATANPERS	AVERR6	AVTOTB	CDCARE	TRABSCOR
	AVERR1	COPYSCOR	BNTTOTAL	CDHOME	TRABSCOR	
	AVERR2	CLOCKHAND	AVERRB	BNTSPONT	CATVGPERS	
	AVTOT3	TRABERRCOM	AVDELERR1	CATVGPERS	CLOCKTIME	
	CATANINTR	AVERR6	CDORIENT	AVERR1	DSPANFLTH	
	COPYTIME	CLOCKSCOR	AVTOT4	BNTSTIM	TRAASCOR	
	TRAAERRCOM	CATVEGESC	CATANIMSC	CATVEGESC		
	AVTOT5	CDORIENT	CLOCKHAND	COPYNUM		

Table 2. Selected attributes for different months

Step 3: Implement different models to the dataset before and after attribute selection, compare the accuracy and verify the selection

We applied different machine learning models from SciKit-Learn and neural nets from various TensorFlow packages to verify our feature selection (see Table 3). We split the data from each month into 70% training and 30% testing sets and compared the accuracies with and without feature selection. From the comparison, we saw that the selected attributes achieved a higher or comparable accuracy in predicting the disease. In the early and late months, such as M06 and M48, the selected attributes performed better than the full entries. In the middle months, like M18, the selected attributes don't meet our expectation. However, we notice that even the full attributes of this month don't give high accuracy. It indicates that the diagnosis recipe currently in use for these stages may not be working ideally.

		Accuracies reported on 30% of the data (used while testing)											
		m6		m12		m18		m24		m36		m48	
	No. of attributes	63	12	63	12	63	12	63	12	63	10	63	5
Models compared	Neural Net	85.38	87.21	85.22	85.71	80.21	82.41	78.88	84.83	87.93	85.34	86.66	83.33
	Decision Tree	82.19	83.1	87.68	82.26	75.82	78.02	80.72	81.32	82.75	82.75	90	93.33
	KNN (k=9)	83.56	84.93	88.66	81.28	84.61	82.41	81.92	77.71	82.75	85.34	73.33	87.32
	SVC	85.38	84.47	91.62	80.78	83.51	80.21	82.53	80.12	86.2	87.06	83.33	86.66
	RandomForestClassifier	86.3	85.84	91.62	81.28	85.71	87.91	85.55	83.73	87.93	84.42	86.67	83.33
	AdaBoostClassifier	83.1	84.93	86.69	82.75	84.61	81.31	84.33	84.33	85.34	83.62	83.33	93.33

Table 3. Comparison of accuracies on the test set with full data/selected attributes

Conclusion

Our motivation stems from the inefficiencies that are present in the current Alzheimer's Diagnosis process. With overwhelming costs and the challenges that come from end-to-end implementation, it is essential to the welfare of the participants that a more streamlined diagnosis is identified. By leveraging present-day machine learning algorithms and the Alzheimer's Disease Neuroimaging Initiative, we aim to simplify this clinical diagnosis. Decision Trees, Random Forest Classifiers, Neural Networks, Support Vector Machines, and Nearest Neighbor Classifiers were all essential in this simplification. Ultimately testing over 800 models through brute force hyper-parameterization, we were able to produce very promising results. While our accuracies steadily hovered in the 80% range, there was not a clear answer to whether or not isolating the models to specific tests was beneficial. In some cases, the selective models outperformed the full models, and vice versa. This suggests that the path to a distinct solution extrapolates from some combination of the models, depending on the phase in question. Another potential solution may be uncovered through incorporating prediction models trained on the provided MRI imaging. Our group members each devoted a great amount of time into this project. Quincia reached out to ADNI and preprocessed the data in MATLAB making it usable in ML models while also drafting our proposal. Jaieu used the packages(Keras, Scikit-Learn) to implement the different models and record the accuracies displayed by each of the classifiers. Jack spent the majority of their time focused on refining our data and presenting results, along with visualizations for the accuracies obtained on various models on different setting and hyperparameters. Ikhlas worked on the tabulation, website design and brought his knowledge and background in neurodegenerative diseases. Our findings will be pivotal in developing an early Alzheimer's diagnosis process and, hopefully shortly after, a treatment program.