Alzheimer's Disease Progression Prediction via ML Models

Ziyi Chen
Jenny Ding
Yanyi Qian
Zhaoyu Qiao
Jinchen Xie

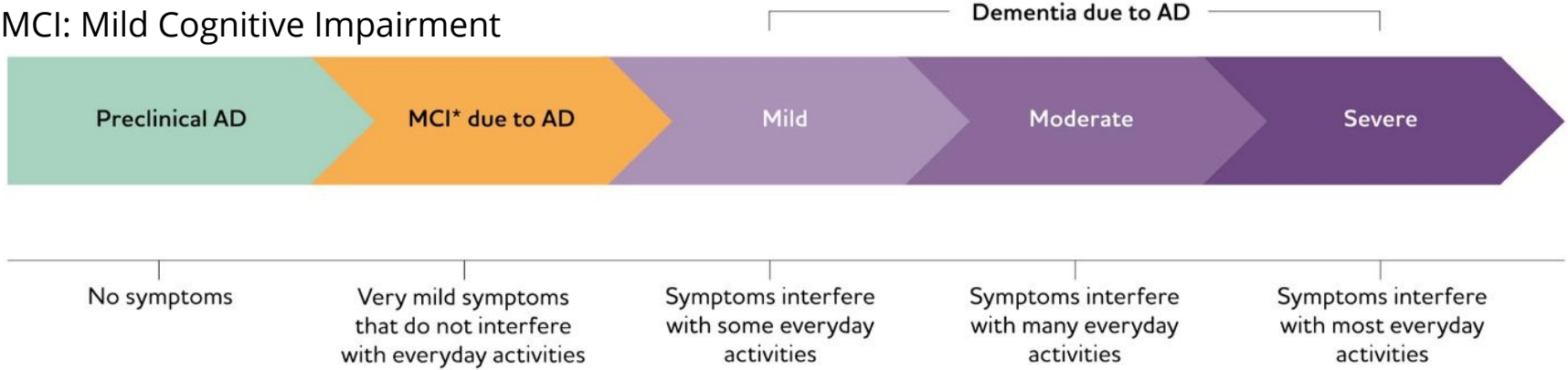
Agenda

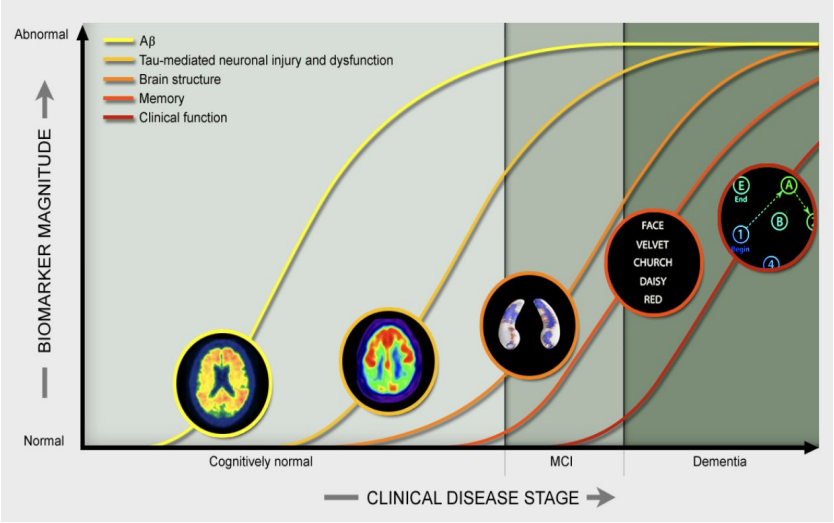
- Background & Goals
- Data description
- Data cleaning
- EDA & Label formation
- Data transformation
- Model & Feature Importance
- R Shiny!

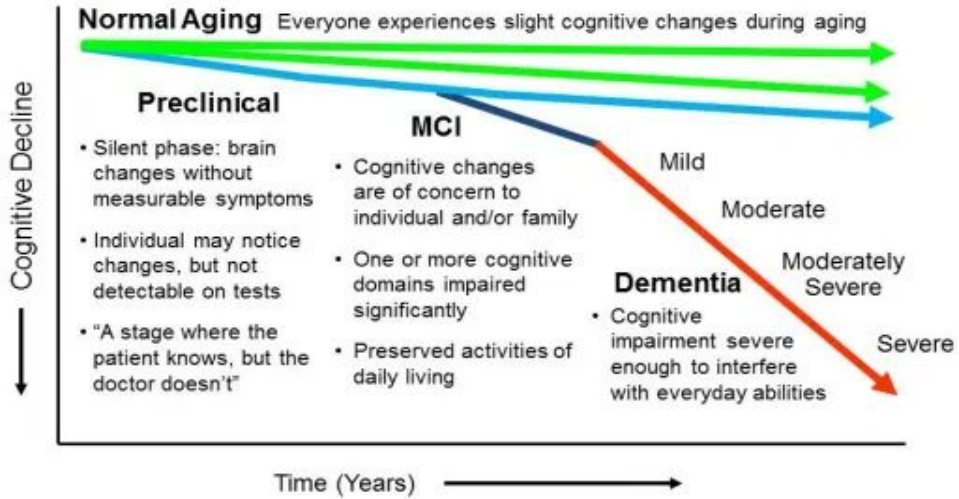
Alzheimer's Disease Facts and Figures

AD: Alzheimer's Disease

MCI: Mild Cognitive Impairment



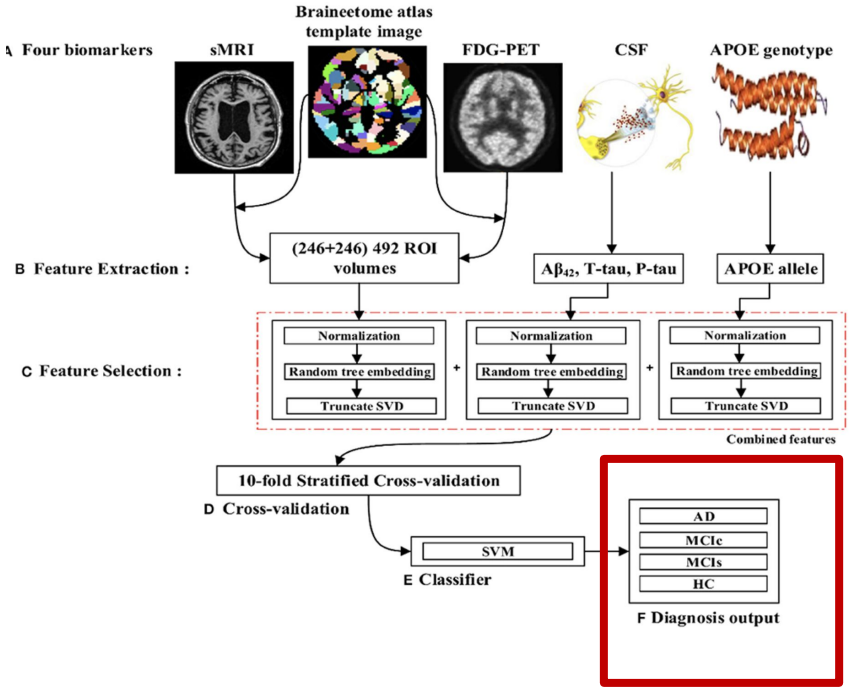




From Full Cognition to Memory Loss

Gaps in Existing Research

- Predicts AD stage rather than transition (model on the right)
  - [Prediction and Classification of Alzheimer's Disease Based on Combined Features From Apolipoprotein-E Genotype, Cerebrospinal Fluid, MR, and FDG-PET Imaging Biomarkers]
- Predicts at a single time point rather than over a period
  - [Predicting Alzheimer's disease progression using multi-modal deep learning approach]
  - [Inexpensive, non-invasive biomarkers predict Alzheimer transition using machine learning analysis of the Alzheimer's Disease Neuroimaging (ADNI) database]
- Lack of a physician friendly UI



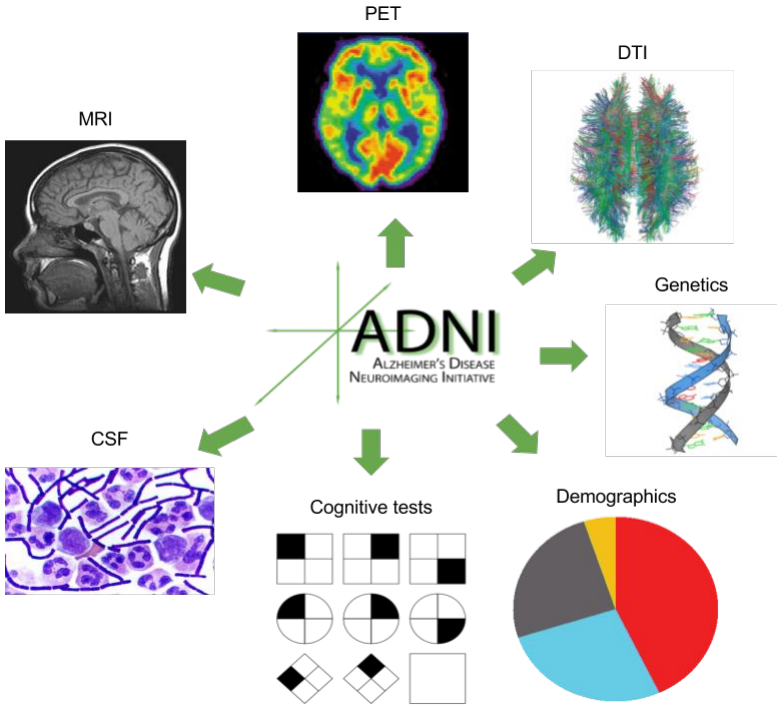
Our goal is to help physicians predict MCI patients' risks of transitioning to AD within 6, 12 and 24 months after the initial visit.

Motivation: “It is essential to improve identification of signs of AD onset because as drugs are developed, early intervention will be an important treatment goal.”

- Sally Temple, Ph.D, co-founder of the Regenerative Research Foundation

ADNI Database and Our Sample

- Alzheimer's Disease Neuroimaging Initiative
- Designed to develop and validate clinical, imaging, genetic and biochemical biomarkers for the early detection and tracking of AD
- Gathered longitudinally from carefully phenotyped subjects by multiple study centers

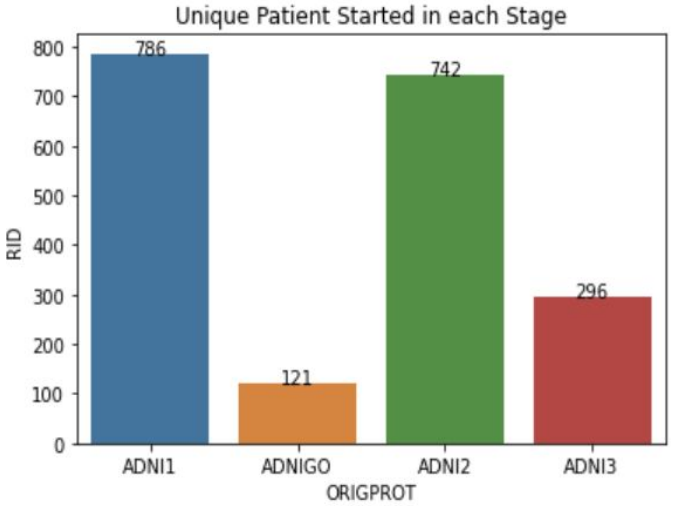


Data Description

Attributes after Data Cleaning

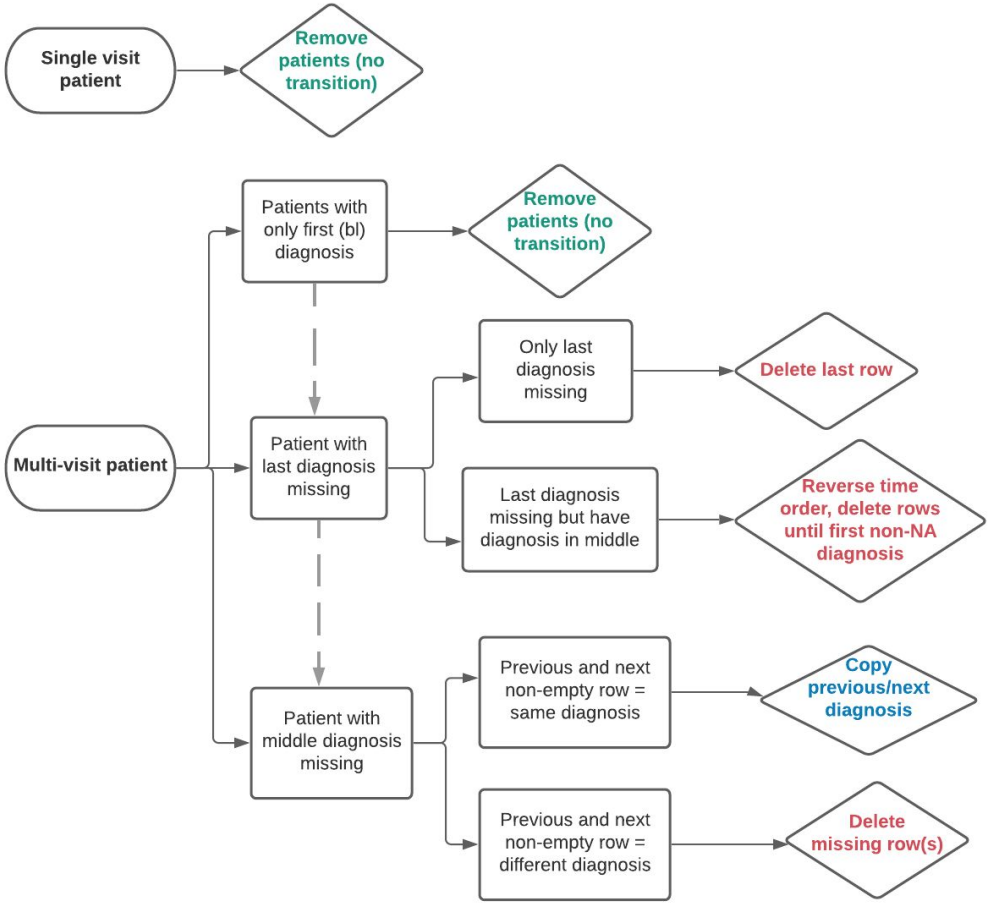
- Biolomarker (3)
  - PTAU, ABETA , etc.
- Demographic (6)
  - Race, marriage, etc.
- Genotype (1)
  - APOE4 allele
- Cognitive test (14)
  - Rey Auditory Verbal Learning Test (RAVLT), Modified Preclinical Alzheimer Cognitive Composite (mPACC), etc.
- Imaging (8)
  - MRI Volumetric Data, PET Imaging

- Duration: 2004-2016
- Size: 1159 patients
- Study stages: 4



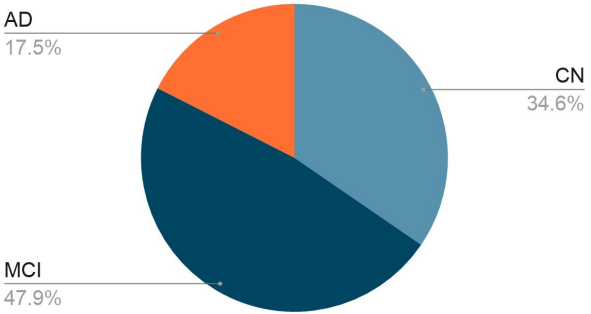
Data Cleaning

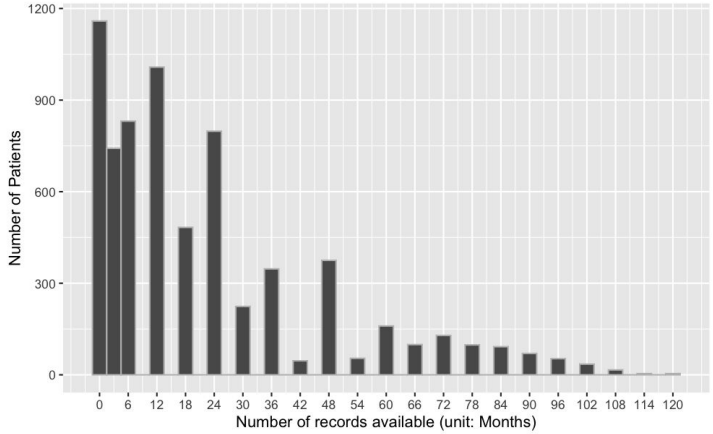
- Feature Removal (>40% missing, meaningless columns)
- Label Imputation (schema on the right)
- KNN Feature Imputation
- One Hot Encoding (label, demographic, APOE4)



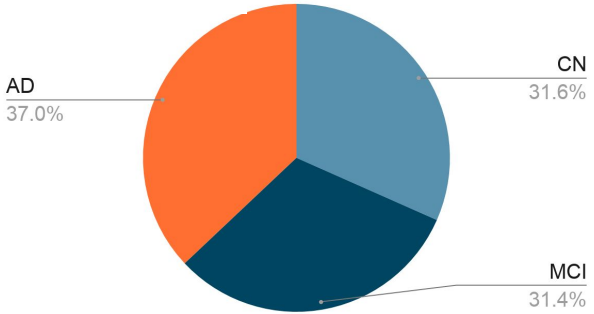
Selecting Early and Late Prediction Time Point

Baseline Diagnosis



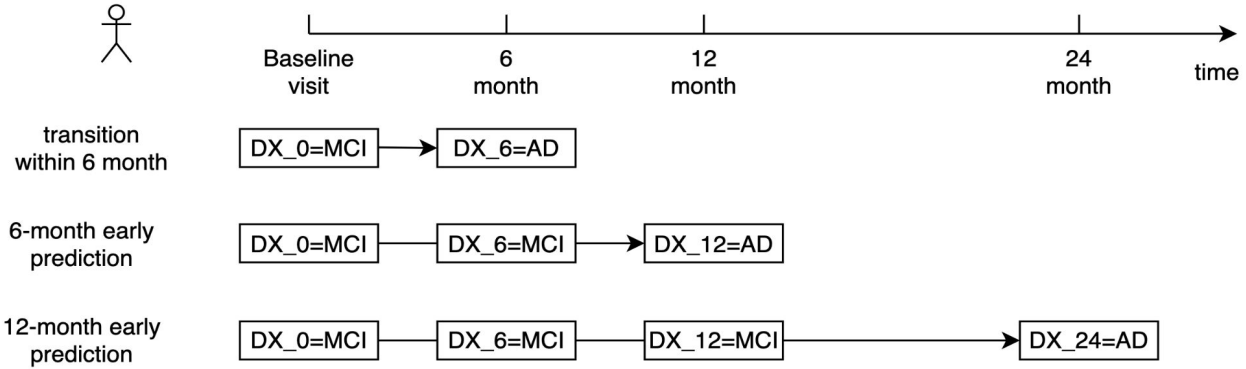


Last Diagnosis



Months Combination	Non-missing	Started with MCI
3, 6, 12	635	370
6, 12, 24	1166	651
12, 24, 36	730	532

Label Formation



Label	Transition within 6 month (n=651)	6 month early prediction (n=617)	12 month early prediction (n=566)
True	34 (5.2%)	51 (8.3%)	95 (16.8%)
False	617 (94.8%)	566 (91.7%)	471 (83.2%)

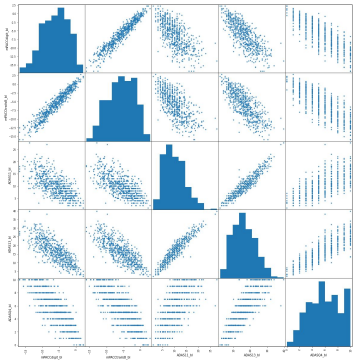
Data Processing

Feature Correlation → **Outlier Detection** → **Data Transfomation**

Pearson Correlation
VIF Statistics

IQR
Isolation Forest

Log Transformation
MinMax Scale

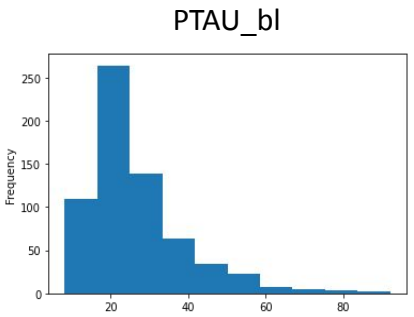


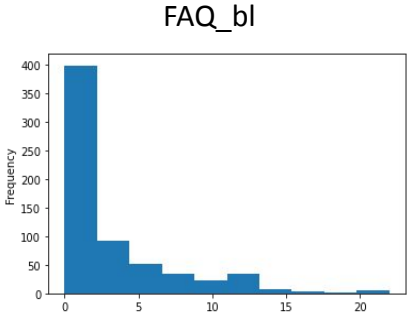
Correlation

mPACCtrailsB & mPACCdigit

PTAU & TAU

ADAS13 & ADAS11





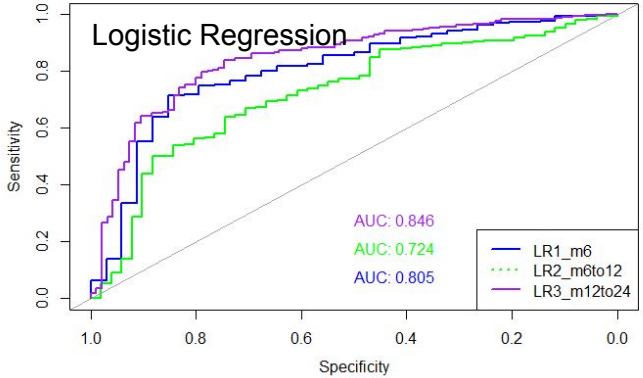
Mean of bl Attributes for Transition & Non-Transition Groups (After 24M)

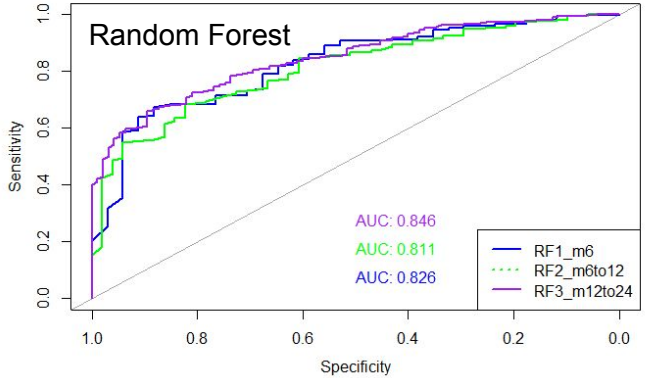
27/41 Attributes w/ p-value < 0.05

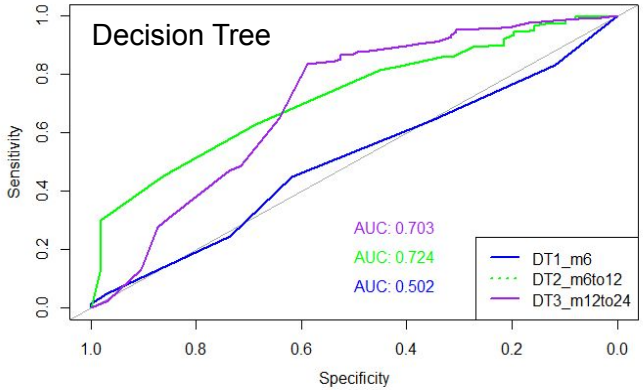
attribute	p_value	no_transition_mean	transition_mean	category
Modified Preclinical Alzheimer Cognitive Composite with Trails test	5.48E-44	0.612	0.376	cognitive
Modified Preclinical Alzheimer Cognitive Composite with Digit test	2.54E-38	0.625	0.402	cognitive
Alzheimer's Disease Assessment Scale -13 Items	1.06E-32	0.384	0.592	cognitive
Alzheimer's Disease Assessment Scale -11 Items	9.60E-29	0.358	0.544	cognitive
Functional Assessment Questionaire	2.54E-28	0.092	0.267	cognitive
Task 4 of the Cognitive Subscale - Constructive Praxis	1.81E-27	0.482	0.719	cognitive
Rey Auditory Verbal Learning Test -Immediate Recall	2.14E-26	0.460	0.289	cognitive
Clinical Dementia Rating - Sum of Boxes	3.07E-22	0.156	0.302	cognitive
Rey Auditory Verbal Learning Test -Precent Forgetting	2.45E-20	0.611	0.829	cognitive
Trail Making Test - Time to Complete	7.79E-19	0.245	0.425	cognitive
MRI Volumetric Data of Hippocampus	8.19E-17	0.537	0.400	imaging
Mini-Mental State Exam	9.94E-17	0.720	0.535	cognitive
Rey Auditory Verbal Learning Test - Rate of Learning	3.35E-14	0.470	0.345	cognitive
Beta-amyloid plaques	**6.51E-14**	**0.532**	**0.349**	**biomarker**
MRI Volumetric Data of Entorhinal	4.58E-13	0.491	0.379	imaging
MRI Volumetric Data of Mid-Temporal Lobe	6.00E-13	0.487	0.384	imaging
MRI Volumetric Data of Fusiform Gyrus	1.69E-10	0.481	0.385	imaging
Number of APOE4 alleles	1.87E-10	0.261	0.455	biomarker
Hyperphosphorylation of tau	1.01E-07	0.204	0.279	biomarker
Neurofibrillary tau tangles	3.83E-07	0.220	0.290	biomarker
MRI Volumetric Data of Whole Brain	3.22E-06	0.418	0.351	imaging
MRI Volumetric Data of CSF Ventricles	4.43E-05	0.227	0.284	imaging
Fluorodeoxyglucose (FDG)-positron emission tomography (PET)	4.52E-05	0.504	0.442	imaging
Rey Auditory Verbal Learning Test - Delayed Recall Score - Immediate Recall Score	0.0038	0.434	0.475	cognitive
Divorced	0.0289	0.104	0.048	demographic
Age	0.0367	0.476	0.514	demographic
Married	0.0461	0.764	0.838	demographic

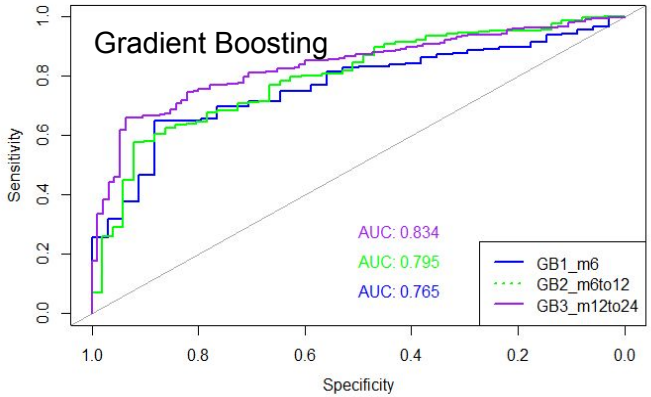
Brief description of each cognitive test: link

Model comparison (ROC CV = 5)



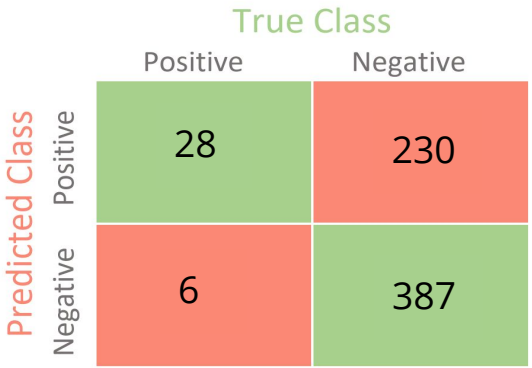






Best Model: Random Forest Confusion Matrix

6M



Sensitivity = TP / (TP + FN) = 0.824
Specificity = TN / (FP + TN) = 0.627

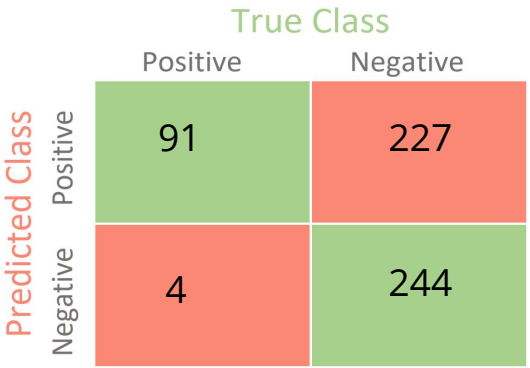
False Negative Rate = FN / (FN + TP) = 0.177
False Positive Rate = FP / (FP + TN) = 0.3728

12M

- We tuned the model to be **HIGHLY SENSITIVE.**
- **False negatives** are tuned to be **MINIMAL**.

- Not predicting a patient might transition is more costly than predicting a patient to transition when he/she doesn't.

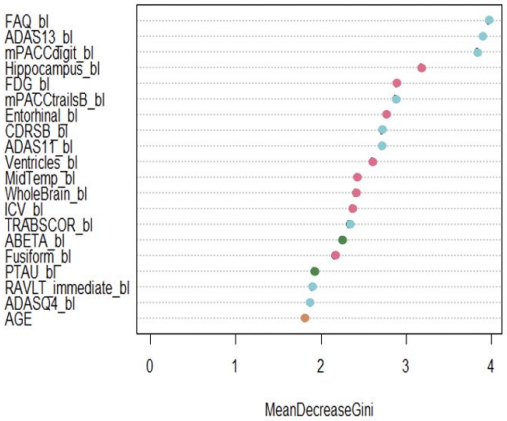
24M



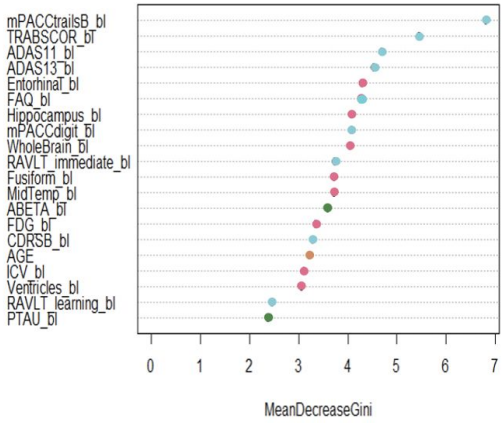
Sensitivity = TP / (TP + FN) = 0.958
Specificity = TN / (FP + TN) = 0.518

False Negative Rate = FN / (FN + TP) = 0.0421
False Positive Rate = FP / (FP + TN) = 0.482

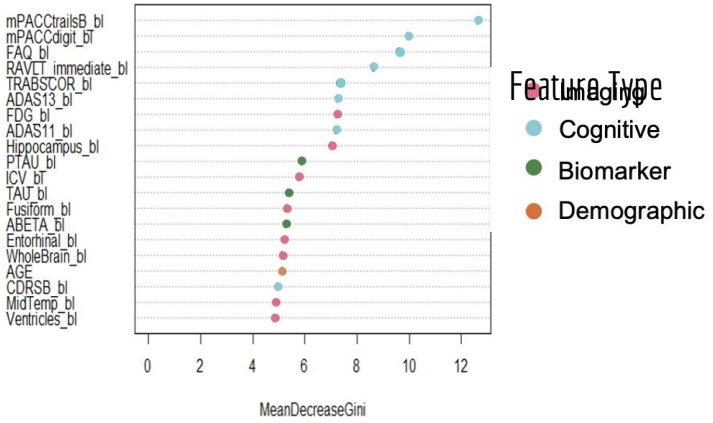
6M Immediate Prediction Feature Importance



12M Intermediate Prediction Feature Importance

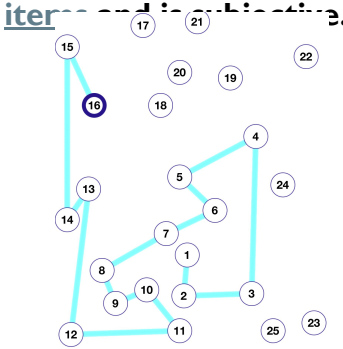


24M Early Prediction Feature Importance

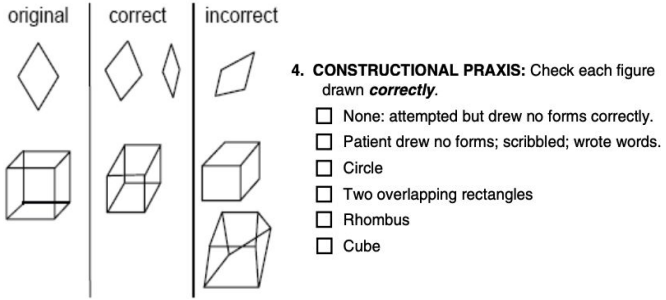


- Cognitive test is the most effective predictor
  - The modified Preclinical Alzheimer Cognitive Composite **(mPACC) digits and trails tests** are most effective for **early prediction**.
  - The Alzheimer's Disease Assessment Scale **(ADAS)** and Functional Activities Questionnaire Index **(FAQ)** are most effective for **immediate prediction.**
- Imaging data's effectiveness drops as prediction time lengthens
  - MRI volumetric data of **hippocampus** and **entorhinal** and PET imaging with F-flurodeoxyglucose **(FDG)** are helpful in **immediate prediction.**
- Biomarkers such as **phosphorylated tau**'s importance rise in predictions for **early prediction**.
- **Age** is the only demographic factor that reached the top 20 important features list.

- **MPACC** Digit and Trail tests are most effective in **early prediction**. **Both are short tests.**
- **ADAS** and **FAQ** are most effective for **immediate prediction.** **ADAS is composed of 13 tests. FAQ has 10 items and is subjective.**



Example of mPACC Trail Test



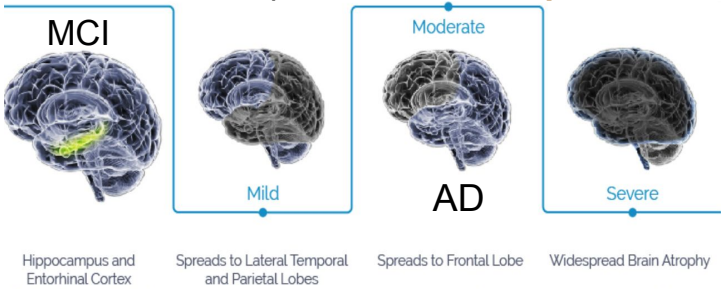
Examples of the 4th item in ADAS

In the Past 4 Weeks, Did the Subject Have Any Difficulty or Need Help With:

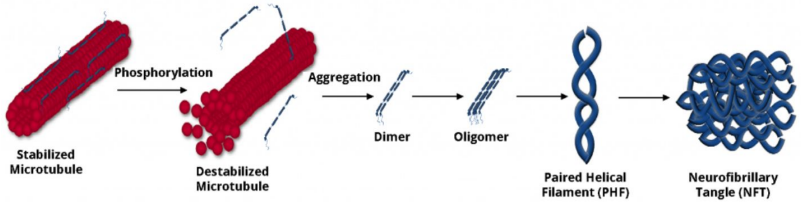
1. Writing checks, paying bills or balancing a checkbook
2. Assembling tax records, business affairs, or other papers
3. Shopping alone for clothes, household necessities, or groceries
4. Playing a game of skill such as bridge or chess, working on a hobby
5. Heating water, making a cup of coffee, turning off the stove

First 5 questions in FAQ

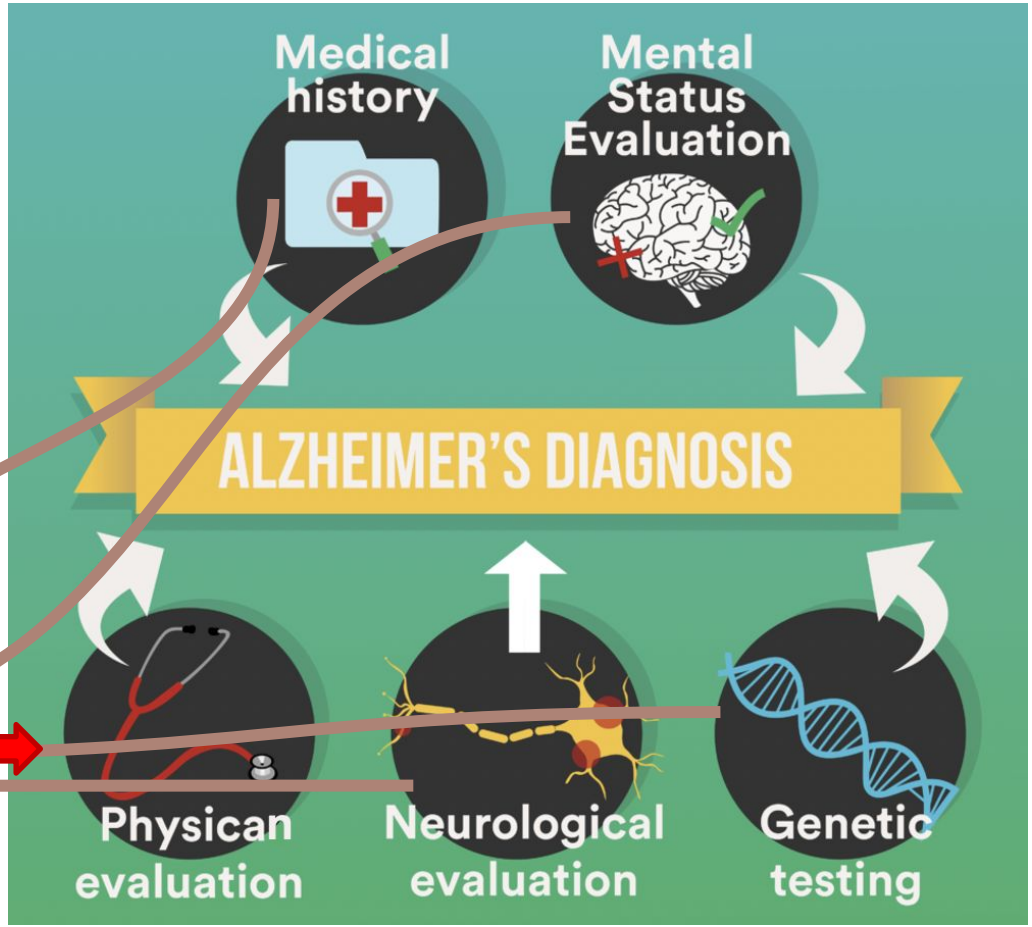
- MRI volumetric data of **hippocampus** and **entorhinal** and FDG PET scan are helpful for **immediate prediction.**



- **Phosphorylated tau**'s level in CSF is important for **early prediction.**



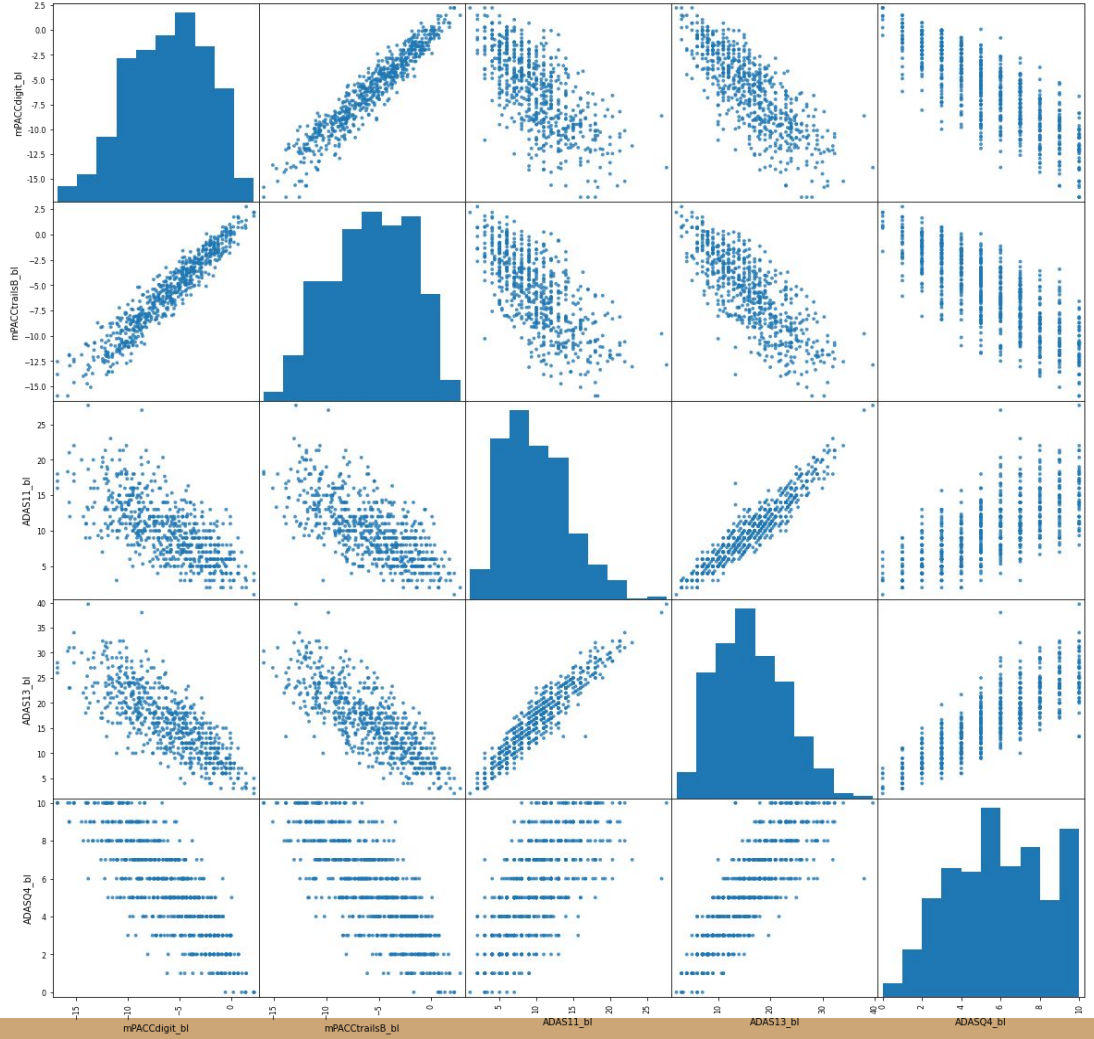
AD Predictor



Appendix

Model comparison (Confusion Matrix)

		Trans_M6		Trans_M6-M12		Trans_M12-M24	
		1	0	1	0	1	0
Random Forest (0.04, 0.07, 0.085)	**1**	28	230	44	196	91	227
	0	6	387	7	370	4	244
Decision Tree (0.03, 0.04, 0.04)	**1**	25	501	40	308	72	282
	0	9	116	11	258	23	189
Logistic Regression (0.001, 0.01, 0.06)	**1**	31	336	46	330	84	164
	0	3	281	5	236	11	307
Gradient Boosting (0.02, 0.03, 0.05)	**1**	30	217	40	201	89	162
	0	4	400	11	365	6	309



Feature Collinearity is Expected and Present (No Removal)

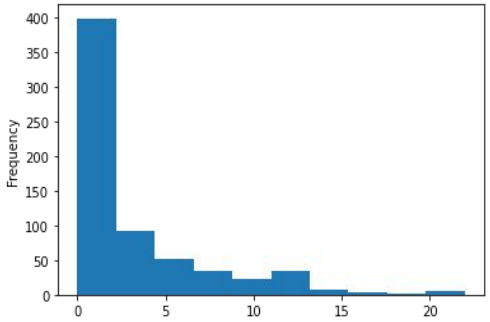
- mPACCtrailsB_bl & mPACCdigit_bl
- PTAU_bl & TAU_bl
- ADAS13_bl & ADAS11_bl

Skewed Data -› Transformation

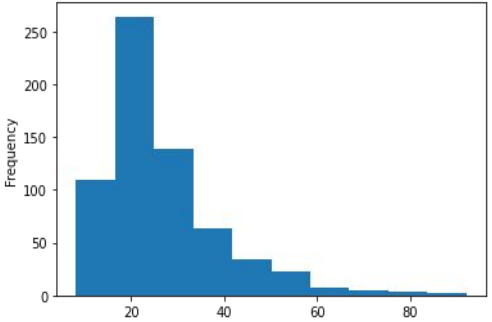
	Value	Mean
FAQ_bl	11.00	3.105991
CDRSB_bl	4.00	1.491551
ADAS11_bl	27.67	10.049631
ADAS13_bl	39.67	16.215499

	Value	Mean
TAU_bl	816.90	270.861925
PTAU_bl	92.08	26.032504
FAQ_bl	11.00	3.105991
CDRSB_bl	4.00	1.491551

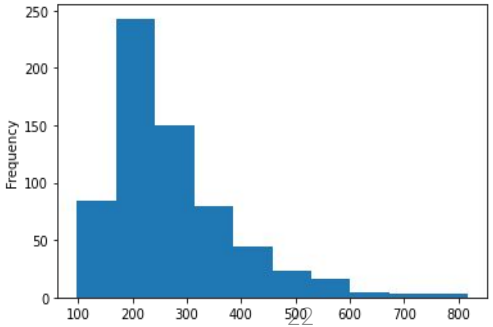
FAQ_bl



PTAU_bl



TAU_bl



Ventricles_bl