

# Exploring the Role of Brain Anatomy in Cognitive Function

---

Rustin Jafarkhani, Brian Jing, Ryan Prabhu, Niyanth  
Raja, Sehaj Sethi, Newton Tsaur, Zhongwen Zhang



# SETTING THE FRAMEWORK

---



## RESEARCH QUESTION

How does **physical** brain anatomy correspond to changes in people's cognition levels?



## CONNECTION

How can the information provided by an MRI scan be used to **detect/predict** Alzheimer's disease in a patient?



## BIGGER PICTURE

How should medical institutions **incorporate** resonance imaging results into their early AD screening process?

---

## HYPOTHESIS

$H_A$  = patterns can be derived from information regarding a patient's brain volume, which can be used to detect Alzheimer's disease earlier than current standards

# DATA SCIENCE VOCABULARY

---

## OVERFITTING

---



## REGRESSION

---



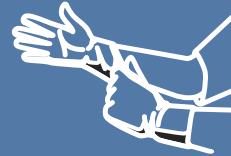
## PREDICTOR

---



## CLASSIFICATION

---



# SCIENTIFIC CONTEXT - ALZHEIMER'S DISEASE

---



1

Right / Left  
Hippocampus

2

Right / Left  
Entorhinal Cortex

3

Cerebellum

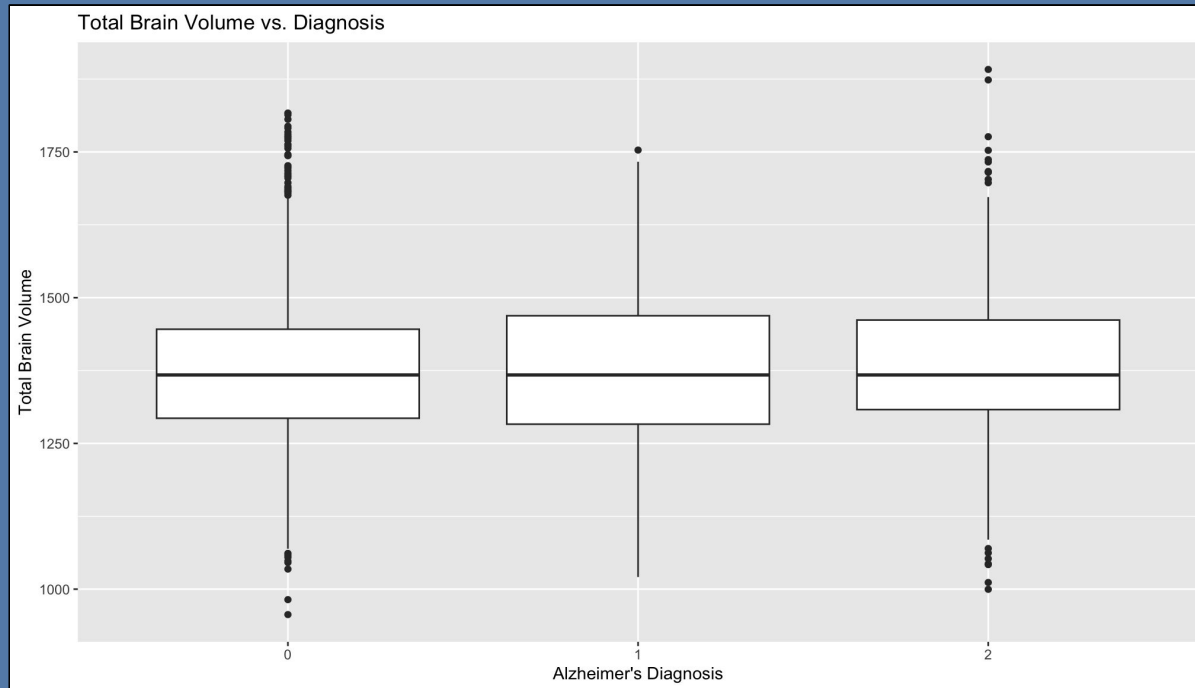
4

Temporal Lobe

# FULL PROCEDURE

1. Using R, install and upload necessary packages (tidyverse, readr, tidymodels, ggplot2, class, dplyr, glmnet, randomForest)
2. Using the readr function, import the alzheimer data set provided by the National Alzheimer Split Coordinating Center
3. Pre-processing Data: Split data into training and testing, this can be done by selecting a random proportion of the alzheimer data set and assigning it to be used as either training or testing data.
4. Using ggplot2, create box-plots, using cerebral spinal fluid volume and brain tissue volume as a function of diagnosis.
  - a. Repeat this process for each subsection of the brain, such that every subsection is a factor of diagnosis.
5. Using the factors which demonstrated a strong correlation with diagnosis, create a multinomial LASSO regression model with the glmnet package, and create a complexity matrix to calculate the accuracy.
  - a. Note which factors the LASSO model determines to be insignificant/drops
6. Using the same factors the LASSO model utilizes, create a random forest model using the randomForest package, with diagnosis as its function.
  - a. Determine the accuracy of the model
    - i. Run the same model, but group the 1, and 2 diagnosis to create a binary response (as opposed to ternary)
7. In order to understand the impact of SMOTE, use the synthetic minority oversampling technique on the data set before splitting it into testing and training data. Therefore, everything will operate with the sampling technique applied.
  - a. Re-evaluate results, and determine if the SMOTE method is worthwhile.

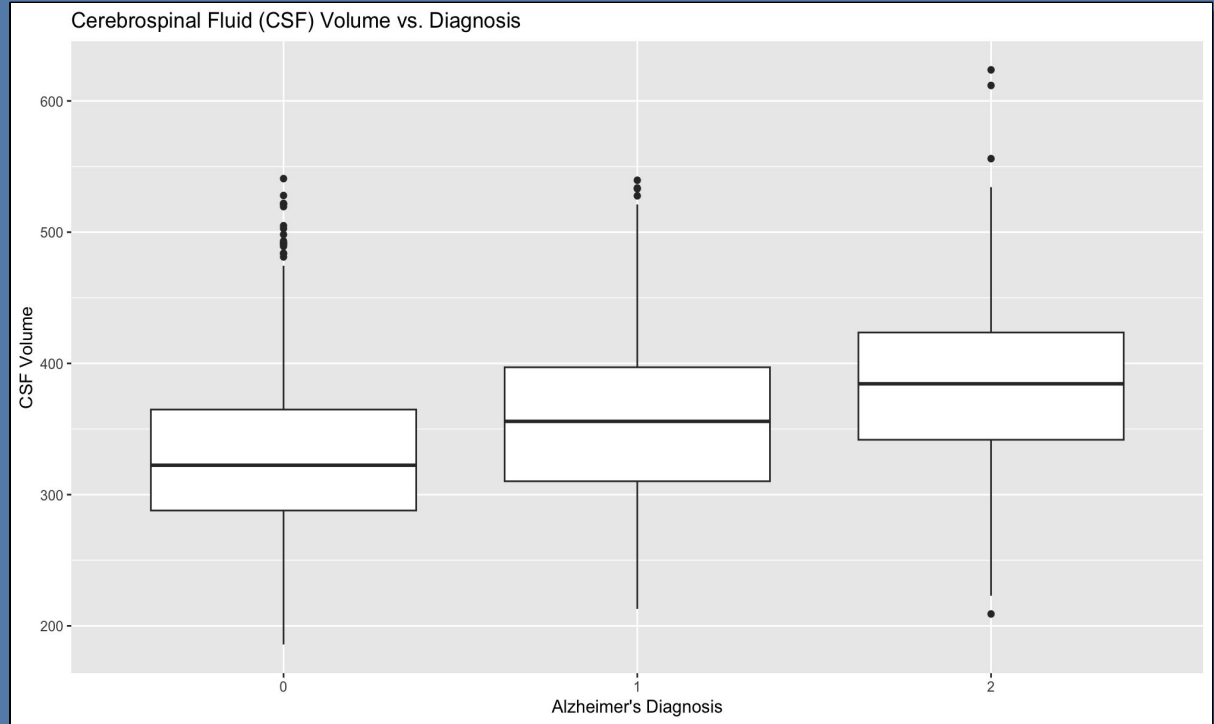
# Total Brain Volume For Each Diagnosis



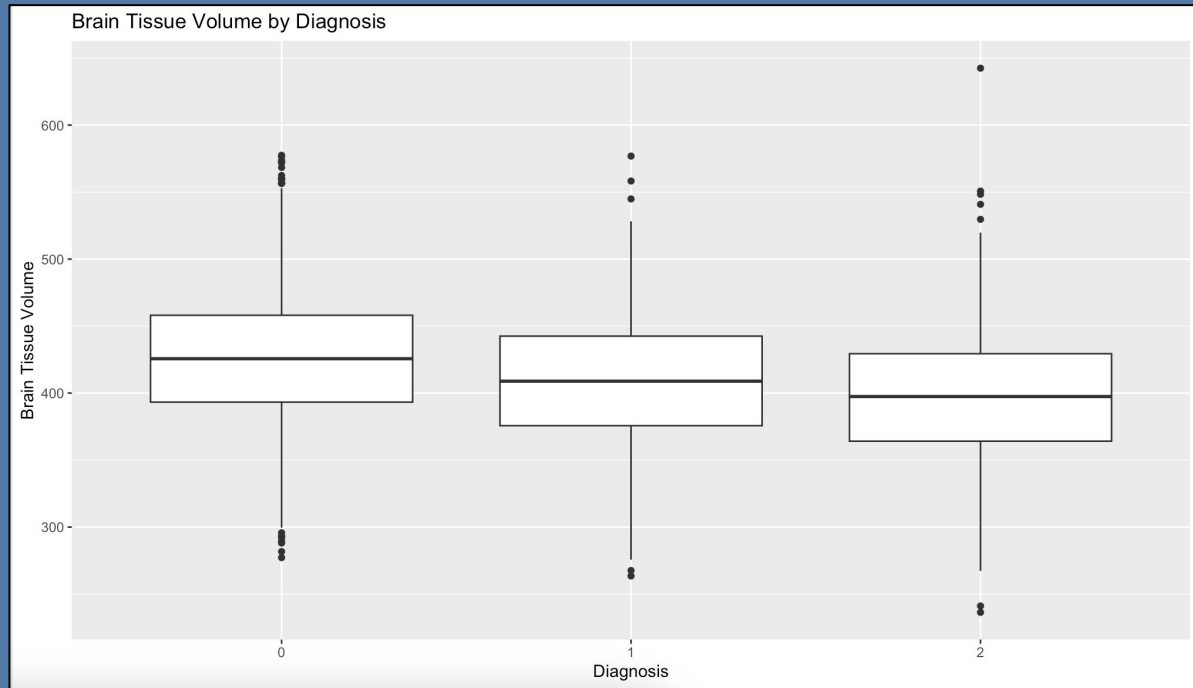
0 = Healthy  
1 = Mild Cognitive Impairment  
2 = Alzheimer's Disease

# Cerebrospinal fluid (CSF) Volume For Each Diagnosis

0 = Healthy  
1 = Mild Cognitive Impairment (MCI)  
2 = Alzheimer's Disease (AD)



# Brain Tissue Volume For Each Diagnosis



0 = Healthy  
1 = Mild Cognitive Impairment  
2 = Alzheimer's Disease



# Multinomial LASSO Regression Model

- Predicts diagnosis status as a function of several brain measurements
- Multinomial model allows for three categorical outcomes
- LASSO is a variable selection technique which eliminates less important variables
- glmnet function in r used for this analysis

$f(x_1, x_2, \dots)$  = Probability of being healthy

$g(x_1, x_2, \dots)$  = probability of MCI

$h(x_1, x_2, \dots)$  = probability of AD

$x_1$   
 $x_2$   
 $x_3$   
 $x_4$   
.  
.  
.

```
$`0`  
15 x 1 sparse Matrix of class "dgCMatrix"  
          s0  
frcort    -3.248782471  
lparcort   0.002541061  
rparcort   0.007528581  
ltempcor   0.018363054  
rtempcor   .  
lhippo     0.808059438  
rhippo     0.493208496  
lcac       -0.052101965  
rcac       -0.385395060  
rent       0.072950148  
lent       .  
rparhip    -0.032507765  
rposcin    .  
lposcin    -0.244716916
```

```
$`1`  
15 x 1 sparse Matrix of class "dgCMatrix"  
          s0  
frcort     0.25653304  
lparcort   -0.01196743  
rparcort   .  
ltempcor   .  
rtempcor   0.04087388  
lhippo     .  
rhippo     .  
lcac       .  
rcac       .  
rent       -0.21557458  
lent       -0.03068504  
rparhip    .  
rposcin    .  
lposcin    .
```

```
$`2`  
15 x 1 sparse Matrix of class "dgCMatrix"  
          s0  
frcort     2.99224943  
lparcort   0.02032823  
rparcort   -0.02720653  
ltempcor   -0.09804183  
rtempcor   -0.02110247  
lhippo     -0.17745331  
rhippo     -0.87821217  
lcac       .  
rcac       0.17948942  
rent       .  
lent       0.37074342  
rparhip    0.54781513  
rposcin    0.05945505  
lposcin    0.01568666
```

# Multinomial LASSO Regression Model

	Chance of healthy	Chance of MCI	Chance of AD diagnosis
	X0 <dbl>	X1 <dbl>	X2 <dbl>
Patient 1	0.39627950	0.32439674	0.279323757
Patient 2	0.76208192	0.16109529	0.076822794
⋮			
Patient 811			

## Assigning by highest probability

811 predictions

**64%**

accurate

## 60% Threshold

443 predictions

**75%**

accurate

## 80% Threshold

173 predictions

**85%**

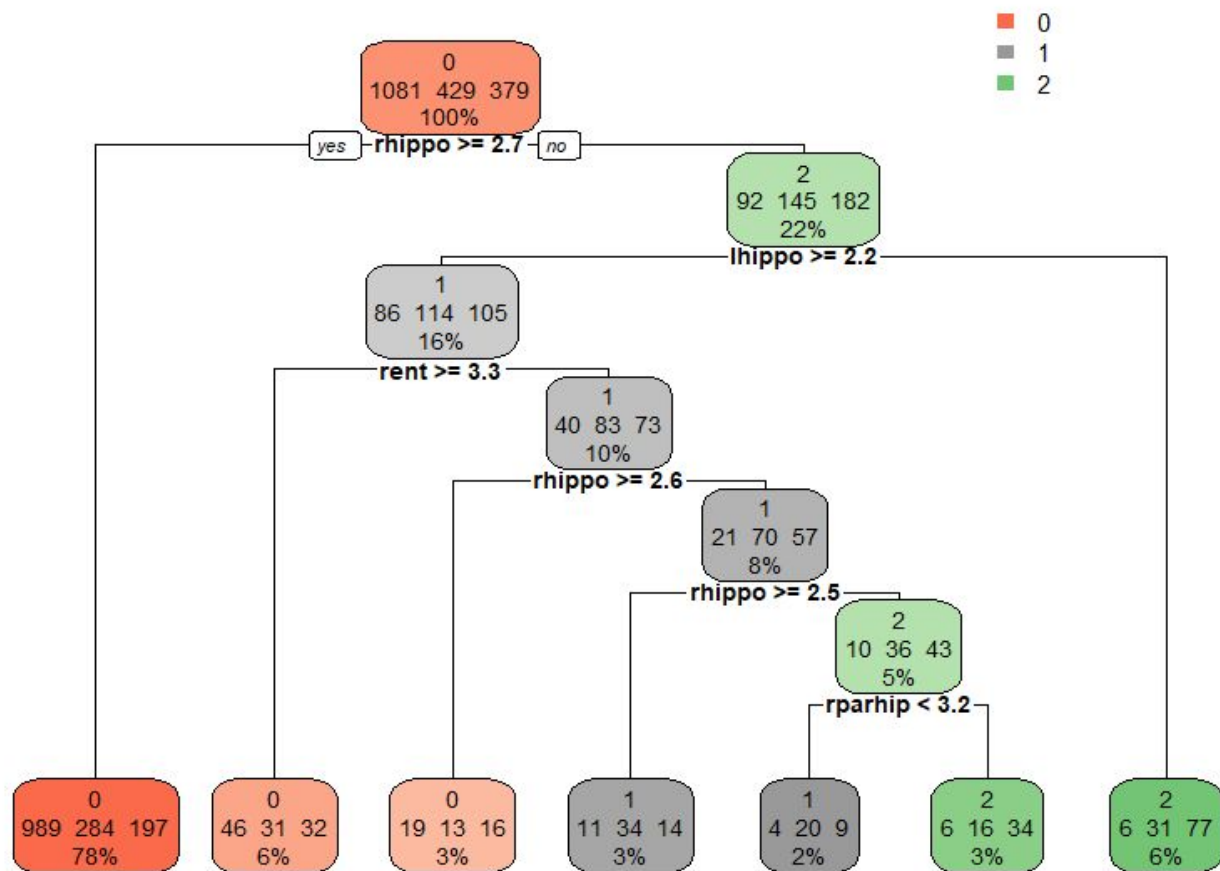
accurate

## 90% Threshold

44 predictions

**93%**

accurate



# Random Forest

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
2	2	0	1	1	2	0	0	0	0	0	2	0	0	0	0	1	1	0	2	2	2	0	2	0	0	0	2	1	1	1	0	0	1	0	1	2	0	0
40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78
0	0	0	2	0	0	0	0	2	0	1	0	0	0	0	1	0	0	0	2	2	0	0	0	0	0	1	0	0	0	0	1	1	0	1	2	0	0	0
79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117
0	0	2	0	0	0	0	0	0	2	0	0	2	0	0	0	0	0	2	0	0	1	2	2	1	0	1	0	0	0	1	0	2	0	1	0	0	0	0
118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156
2	0	0	0	1	1	0	0	1	0	2	0	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0
157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195
1	1	0	1	0	0	0	0	0	2	2	2	0	1	2	0	0	0	0	0	0	0	2	1	0	2	0	1	0	2	2	0	2	0	0	0	0	0	0
196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234
0	0	0	0	0	0	2	1	1	0	0	0	2	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0
235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273
2	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	
274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312
0	0	0	2	2	0	0	0	0	0	0	1	0	0	0	0	0	1	2	0	0	0	0	1	2	0	0	1	0	0	0	0	0	0	0	2	0	0	1
313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351
0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0
352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390
0	0	0	0	1	0	0	0	1	0	1	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	1	0	0
391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429
0	0	0	2	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	2	2	0	0	0	0	0	2	2	0	0	0

Binary: 90.38%

Ternary: 84.83%

# Using SMOTE (Synthetic Minority Oversampling Technique)

0	1	2
1534	613	553



Current Dataset With Diagnosis  
Classes (no upsampling)

0	1	2
0.5722605	0.2271043	0.2006353



Percentages In Terms Of Total  
Patients

0	1	2
4017	3840	3663



After applying SMOTE upsampling  
technique

# Discussion - Multinomial LASSO



How should medical institutions **incorporate** resonance imaging results into their early AD screening process?

Multinomial LASSO Regression Model

X0 <dbl>	X1 <dbl>	X2 <dbl>
0.39627950	0.32439674	0.279323757
0.76208192	0.16109529	0.076822794

60% threshold; 75% accurate predictions

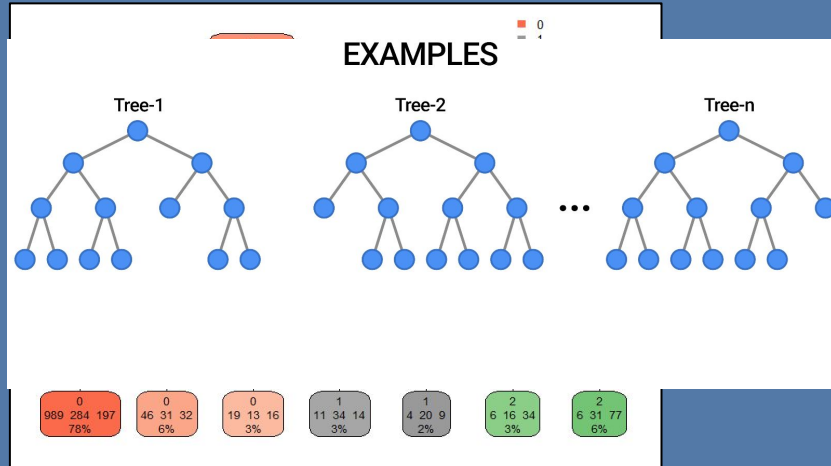
80% threshold, 85% accurate predictions

90% threshold, 93% accurate predictions

What do these values mean? How do they help us answer our focus question?

- Threshold use cases
  - Hospital/Diagnosis use
    - Priority-lists/pre-screening
    - Second Opinion

# Discussion - Random Forest



Random Forest Classification Model Results

**Binary: 90.38% Ternary: 84.83%**

What do these values mean? How do they help us answer our focus question?

- Use in conjunction with original model
  - Hospital/Diagnosis use
    - Priority-lists/pre-screening
    - Second Opinion

**Moving Forward: How  
would we continue our  
research?**





# THANK YOU!

CREDITS: This presentation was made possible with the extensive assistance and knowledge of Jonathan Duya, Dr. Magda El Zarki, Lynn Gao, Dr. Jessica Jaynes, Julie Lindo, Brian Schetzslé, Dr. Babak Shahbaba, and Dr. Michael Shindler.