

# Aging, Dementia and TBI Study

Himank Kansal, Sudeep Singh, Chen Xu

## 1. Introduction

With an aging population that continues to grow, we are seeing more and more people having neurodegenerative diseases, such as Dementia and Alzheimer's disease. These diseases are caused by damage to nerve cells (neurons) in the brain. About 11% of the US population over the age of 65 are believed to be affected by dementia. The incidence of dementia onset is estimated to double every five years to about 40% to 60% after age 90. This significantly increases demands on the public health system and on medical and social services, causing large economic burden and healthcare challenges.

As a group of data scientists who want to analyze real world data to get useful insights, we have worked on datasets from the Adult Changes in Thought (ACT) study (<http://aging.brain-map.org/overview/home>). The ACT study is a longitudinal population-based prospective cohort study of brain aging and incident dementia in the Seattle metropolitan area. We are interested in understanding the relationship between cognition, brain pathology and injury in the aged brain.

We analyzed the donors' de-identified clinical information and tissue samples' metrics from different imaging technologies such as Histology and immunohistochemistry (IHC) and Luminex. We built a supervised learning model that predicts whether a person has dementia or no dementia using demographic information and certain metrics from tests. We also drew useful insights from the visualization we built.

## 2. Related Work

As the ACT study was supported by the Paul G. Allen Family Foundation, much research work has already been done. The website <http://aging.brain-map.org/overview/explore> documents many insightful findings and includes cool interactive visualizations for readers to explore the relationship between dementia and protein levels, top genes identified in the research.

We found most of the scientific findings and visualizations on the website were hard to understand for people with little medical knowledge. We were motivated to educate the public about these useful scientific findings with simplified language and intuitive visualizations. We also believe that a predictive machine learning model which can provide users a probability of getting dementia based on metrics from medical tests and exams would arouse many interests of the public about these neurodegenerative diseases.

## 3. Experimental Setup

### 3.1 Data Overview

There are a total of 107 participants (63 males and 44 females) with a wide range of educational backgrounds. The cohort is quite old, ranging from 77 to 102 years old at time of death. Median age is 90. (Figure 1)

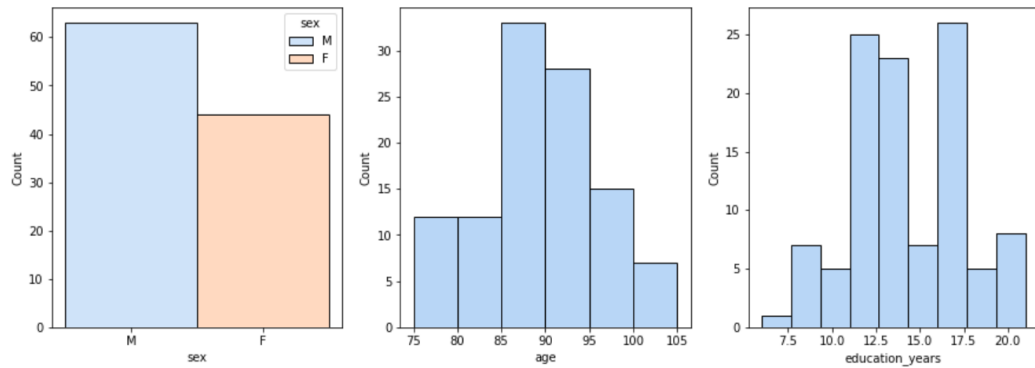


Figure 1

Around half of the donors were diagnosed with dementia. According to DSM-IV (Diagnostic and statistical manual of mental disorders) clinical diagnosis, 30 are diagnosed with AD, 12 with dementia of multiple etiologies, and four with vascular dementia. (Figure 2)

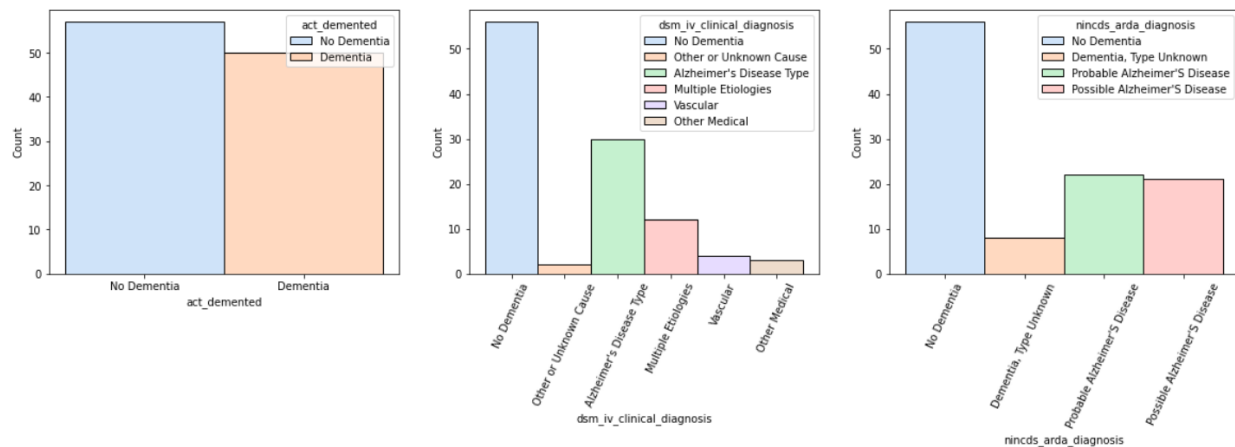


Figure 2

Dementia is an overall term for a particular group of symptoms. The characteristic symptoms of dementia are difficulties with memory, language, problem-solving and other thinking skills. Dementia has several causes that reflect specific changes in the brain: Alzheimer's disease, Parkinson's disease dementia and Lewy body disease.

Approximately 2/3 of dementia cases are diagnosed as Alzheimer's disease (AD), estimated to reach 13.8 million cases by 2050. The brain changes of Alzheimer's disease include the accumulation of the abnormal protein beta-amyloid ( $A\beta$  plaques) outside neurons and twisted strands of the protein tau (tangles) inside neurons in the brain. These changes are accompanied by the death of neurons and damage to brain tissue.

For each donor, tissues were collected from four brain regions known to show neurodegeneration and pathology because of AD and Lewy body disease. They are frontal white matter (FWM), temporal cortex (TCx), parietal cortex (PCx), hippocampus(HIP).

The cerebrum represents one of the largest regions of the brain. The cerebral cortex serves as the outer layer of the cerebrum and it consists of mostly of gray matter whereas white matter lies in the center. Four lobes make up the cerebral cortex: the frontal lobe, the parietal lobe, the temporal lobe, and the occipital lobe (Figure 3). Each lobe has a distinct function.

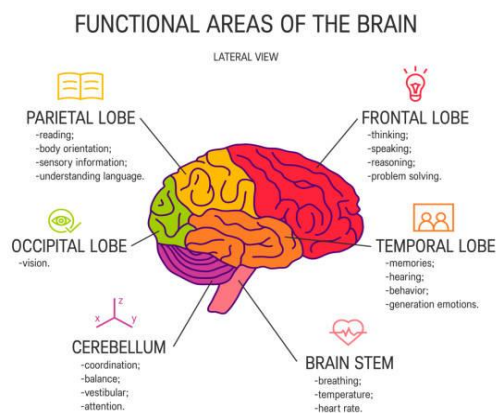


Figure 3

Immunohistochemistry (IHC) was used on both fresh frozen and formalin fixed paraffin embedded (FFPE) tissue to stain and quantify proteins marking dementia-related pathologic findings, including pTau, A $\beta$ ,  $\alpha$ -synuclein (Lewy bodies), and pTDP-43, as well as microglia (IBA1) and astrocytes (GFAP).

Immunohistochemistry (IHC) is an imaging technique to visualize antigens in cells. Labeled antibodies bind to target antigens in the cell to image the distribution and localization of specific proteins of interest.

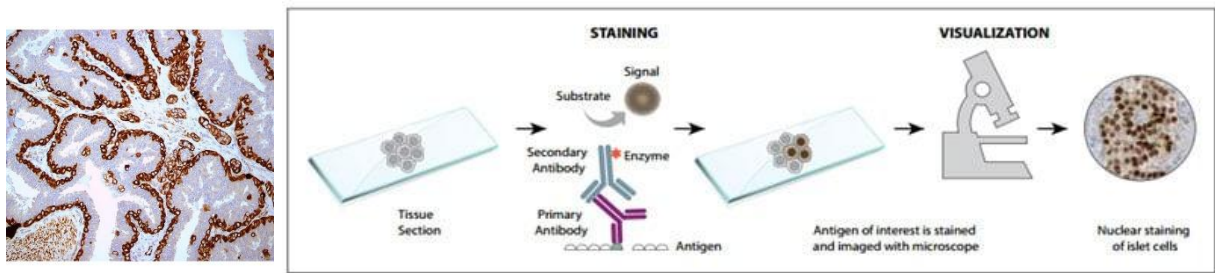
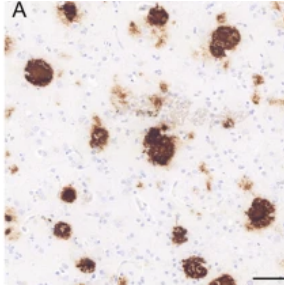
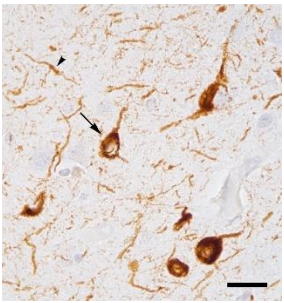
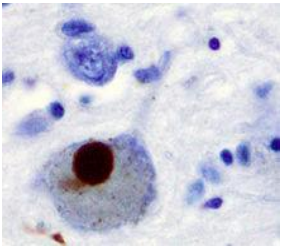


Figure 4

Here is a table which provides more details about the proteins we are interested in.

protein	description	measure	image	location
---------	-------------	---------	-------	----------

beta-amyloid (A $\beta$ )	<p>A<math>\beta</math> is formed from the breakdown of a larger protein, called amyloid precursor protein (APP) and is toxic. Abnormal levels of this naturally occurring protein clump together to form plaques that collect between neurons and disrupt cell function.</p> <p>Cerebrospinal fluid (CSF) levels of amyloid-beta 42(A<math>\beta</math> 42) serve as an excellent marker for brain amyloid as detected by the amyloid tracer, Pittsburgh compound B (PIB).</p>	CERAD (Consortium to Establish a Registry for Alzheimer's Disease) score: 0 (none) to 3 (severe)	 <p>A<math>\beta</math> plaques in an AD patient with an autosomal dominant pattern (ADAD) caused by mutations in the amyloid precursor protein (APP) or in the two presenilin (PSEN1 and PSEN2) genes.</p>	between neurons
pTau	<p>Tau normally binds to microtubules and assists with their self-assembly, formation and stabilization. However, when tau is hyperphosphorylated, it is unable to bind, and the microtubules become unstable and begin disintegrating.</p> <p>The unbound tau clumps together in formations called neurofibrillary tangles (NFT).</p>	Braak stage: 0 (none) to 6 (extensive neocortical NFTs)	 <p>Abnormal accumulation of tau protein, which constitutes NFT, in neuronal cell bodies (arrow) and neuronal extensions (arrowhead) in the neocortex of a patient who died with AD at Braak stage 6</p>	inside neurons
alpha-synuclein ( $\alpha$ -synuclein)	Alpha-synuclein aggregates forming Lewy bodies, a spherical masses in the cytoplasm that displace other cell components, restricting the mobility of synaptic vesicles, consequently attenuating synaptic vesicle recycling and neurotransmitter release		 <p>Positive Alpha-Synuclein staining of a Lewy</p>	inside neurons

			body in a patient with Parkinson's disease	
--	--	--	--	--

Braak Staging is the method used to classify the degree of pathology in Parkinson's and Alzheimer's disease. For AD patient, Braak stages 1 and 2 are used when neurofibrillary tangle involvement is confined mainly to the entorhinal cortex. Stages 3 and 4 are used when there is also involvement of limbic regions such as the hippocampus, and Stages 5 and 6 are used when there is extensive neocortical involvement. In the following image (Figure 5). Yellow represents the origin of Parkinson's pathology. Pink/purple represent Stages 1 and 2. Blue represents Stages 3 and 4. Orange represents Stage 5. Yellow represents full neocortex engagement and Stage 6.

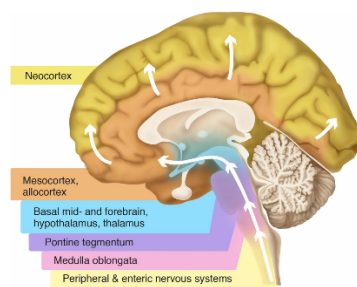


Figure 5

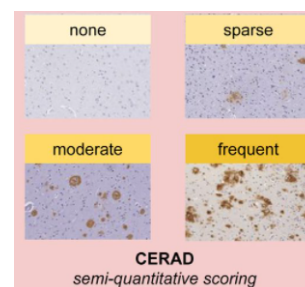


Figure 6

CERAD score is a measure of neurotic plaques which is on a 4-point scale (0 to 3): normal, mild, moderate, severe impairment based on semiquantitative estimates of neurotic plaque density as recommended by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Figure 6 shows the IHC stained plaques.

### 3.2 Preliminary Analysis

We first performed data analysis on the donor information. The following table shows the summary of demographics for donors with and without dementia.

Attribute	Dementia		No Dementia		p-value
	Mean	SD	Mean	SD	
Age at death (years)	90.3	6.5	88.6	7.0	0.21
Education (years)	13.5	3.2	14.7	3.2	0.07
Number of TBIs	0.6	0.7	0.6	0.7	0.87
Age at first TBI	21.1	30.2	20.6	30.7	0.94
Braak stage	4.1	1.7	2.8	1.5	8.7E-05
NIA Reagan	1.9	0.9	1.4	0.7	1.2E-03
CERAD score	1.8	1.2	1.2	0.9	0.01
	Count		Count		
APOE ε4 alleles	13 Yes/32 No		7 Yes/48 No		
Gender	27 M/23 F		36 M/21 F		

Next, we performed data analysis on the protein and pathology information. These metrics for tissue specimens were collected using traditional antibody based IHC methods complemented by Luminex quantification and isoprostane quantification.

Tau protein have roles primarily in maintaining the stability of microtubules in axons and are abundant in the neurons of the central nervous system (CNS). However, when tau is hyperphosphorylated, it is called pTau which is unable to bind, and the microtubules become unstable and begin disintegrating. The unbound tau clumps together in formations called neurofibrillary tangles (NFT).

Phosphorylation-dependent anti-tau antibodies such as monoclonal antibody AT8 is widely used to recognize pTau. We can see from our dataset that the AT8 IHC and pTau Luminex level are strongly correlated (Pearson correlation coefficient is 0.68).

We calculated Pearson correlation coefficient between all the proteins and pathology qualifications and all the attributes of donors. We found that the following correlations are moderate and strong ( $r > 0.4$ ).

Donor attribute	Protein/pathology quantification	correlation
CERAD	ihc_a_beta	0.43
CERAD	ihc_at8_ffpe	0.41
CERAD	ab42_pg_per_mg	0.63
BRAAK	ihc_at8_ffpe	0.46
BRAAK	ab42_pg_per_mg	0.46
NIA Reagan	ihc_at8_ffpe	0.44
NIA Reagan	ihc_a_beta	0.4
NIA Reagan	ab42_pg_per_mg	0.58

It seems that AT8 IHC, A $\beta$  IHC and A $\beta$ 42 are moderate correlated with all CERAD score, Braak Stage and NIA Reagan. As criteria for NIA Reagan diagnosis relies on both neurofibrillary tangles (Braak) and neurotic plaques (CERAD), we didn't include RIA Reagan in our analysis. We focused our analysis on protein molecular quantification of A $\beta$  and pTau.

## 4. Results

### 4.1 Exploratory Data Analysis

From the preliminary data analysis, we found the mean values of Braak Stage and CERAD scores for dementia donors and no-dementia donors are very different. Let's take a closer look at the relationship between different demographics information. Figure 7 shows the boxplots for Age and CERAD score, Age and Braak Stage. It is obvious that older donors are likely to have higher CERAD score and Braak stage.

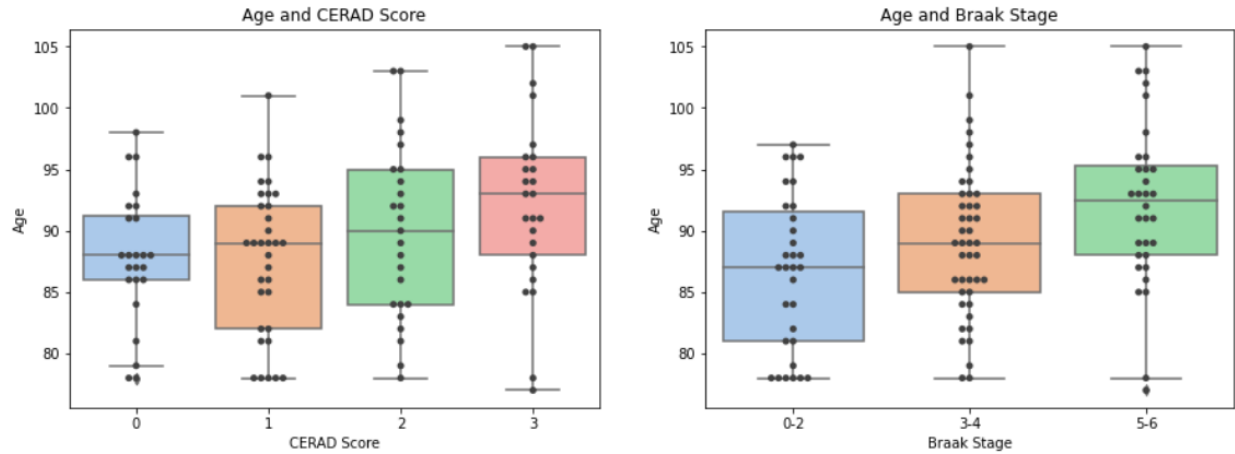


Figure 7

CERAD score and Braak stage are generally higher in donors with dementia compared to donors without dementia, as expected (Figure 8).

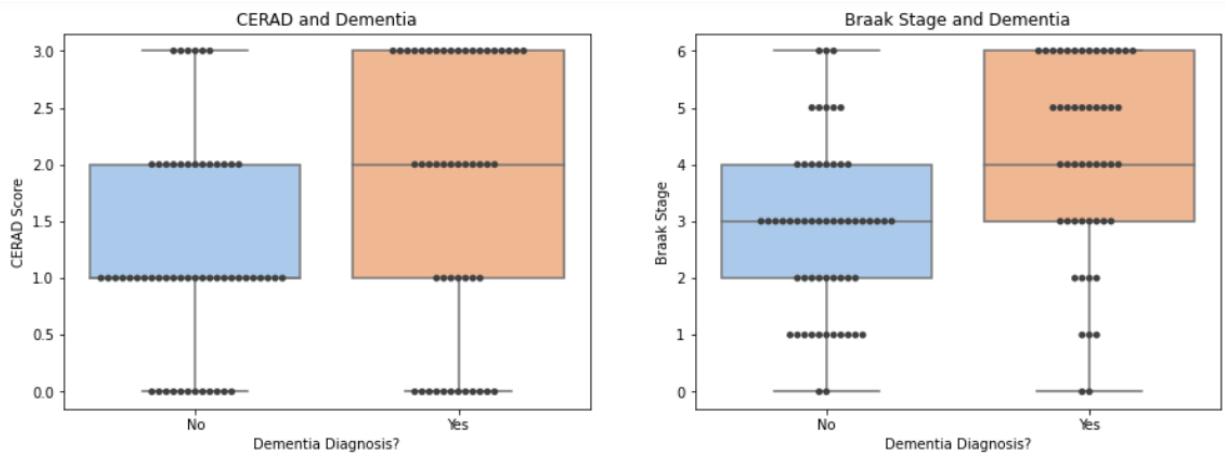


Figure 8

The following Sankey diagram (Figure 9) depicts the relationship between brain region, Braak stage and dementia status. The left column shows the four brain regions, the middle column shows the 7 Braak stage number. The right column shows the two dementia status. The width of each arrow between columns and the height of each bar of the column are based on the quantity. We can observe that samples are evenly distributed among 4 brain regions. Very few samples have Braak stage of 0. More samples have Braak stage of 3 than any other stage. Most samples with Braak stage of 5 or 6 have dementia and most samples with Braak stage of 1, 2 or 3 have no dementia. Almost half of the samples with Braak stage of 4 have dementia. It shows that there is some correlation between Braak Stage and dementia status.

Sankey Diagram: Braak stage and Dementia Status

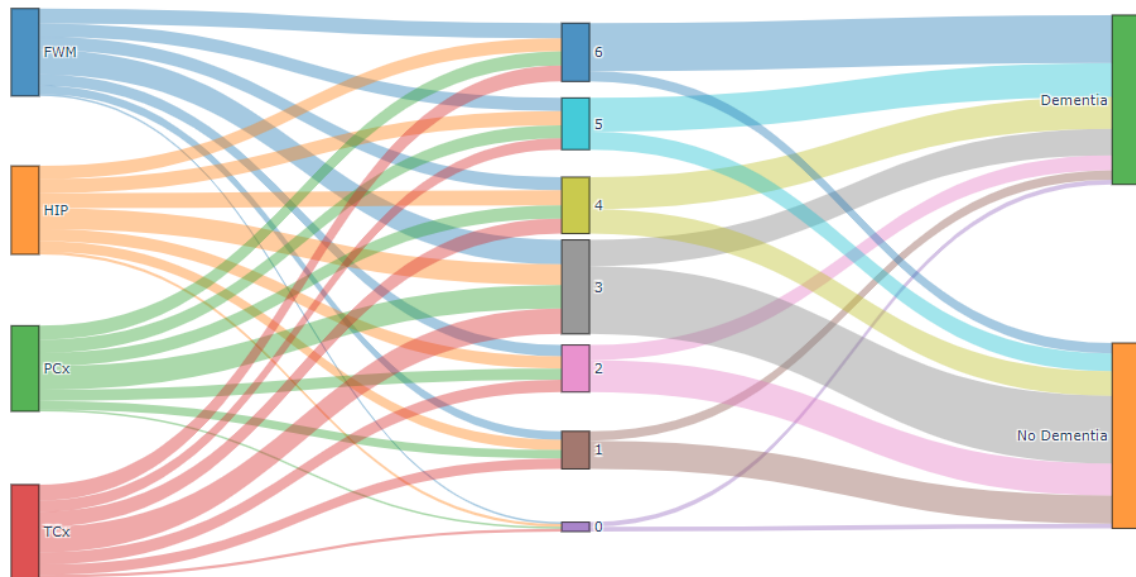


Figure 9 Braak Stage and Dementia Status

Our dataset includes 53 participants reported 1-3 lifetime TBI (Traumatic Brain Injury) with loss of consciousness, along with 54 individuals matched for age, sex, and year of death who did not report a TBI with loss of consciousness.

The following is another Sankey diagram (Figure 10) that depicts the relationship between brain region, number of TBI and dementia status. As the samples with no TBI or 1 TBI are almost evenly distributed between the two-dementia status, we don't really see any strong correlation between the number of TBI and dementia status.



Sankey Diagram: number of TBI and Dementia Status

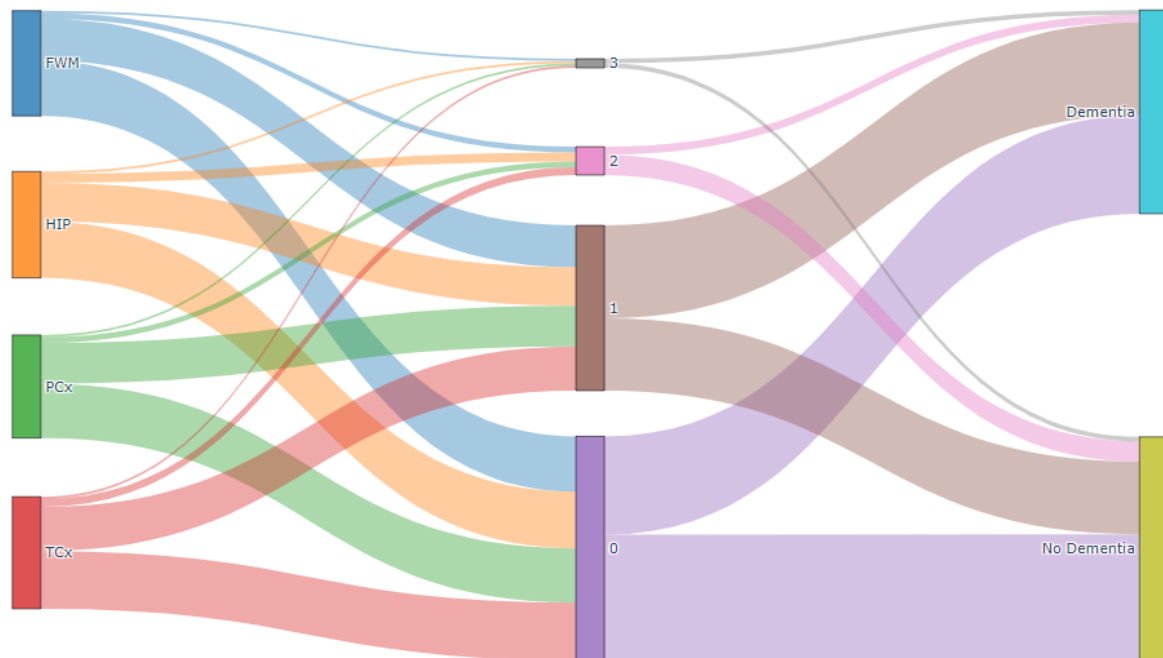


Figure 10

65% of the *APOE*  $\epsilon 4$ -positive participants had dementia while 40% of the *APOE* $\epsilon 4$  negative participants had dementia, consistent with the role of this gene as a strong risk factor for AD. Inheriting *APOE*4 does not mean a person will develop the disease, it does increase the risk for dementia. The *APOE* protein helps carry cholesterol and other types of fat in the bloodstream. Recent studies suggest that problems with brain cells' ability to process fats, or lipids, may play a key role in Alzheimer's and related diseases.

Next, we explored  $A\beta$  and pTau correlations in different brain regions and different Braak stages. Some research studies suggest that formation of plaques and tau phosphorylation might be linked to each other. We analyzed the correlation of  $A\beta$  and pTau for the whole dataset and each brain region. From Figure 11, we saw that  $A\beta$  and pTau have weak correlation (Pearson correlation coefficient is 0.33) overall for all brain regions. But they have moderate correlations in hippocampus and cortex while having no or very weak correlation in frontal white matters. Also, the correlations in later Braak stages (5 and 6) are much stronger than early stages (0-4).

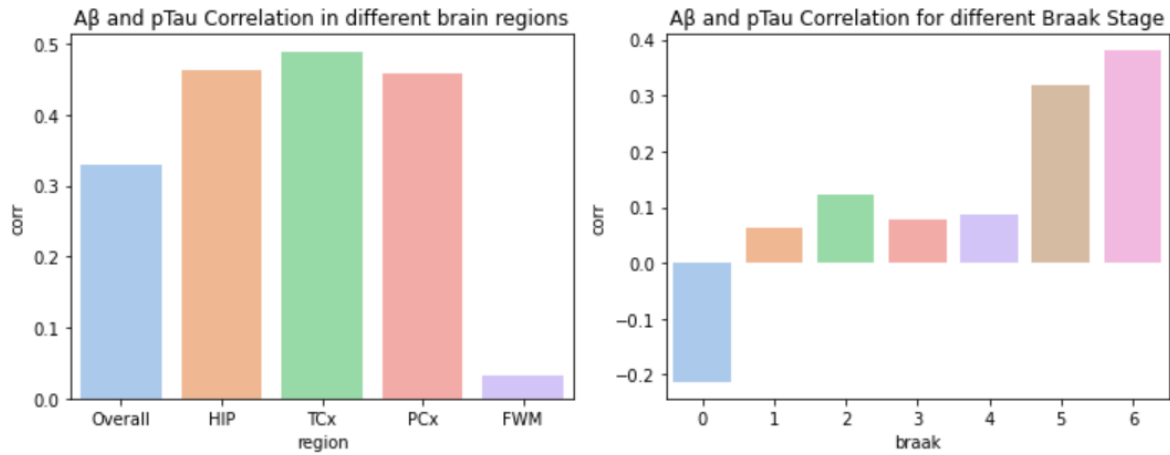


Figure 11

Our analysis reveals pathological tau (AT8 IHC and pTau Luminex) tended to be higher in hippocampus while Aβ (Aβ IHC and Aβ42 Luminex) is higher in cortex as shown in Figure 12, consistent with known AD pathological distributions and progression.

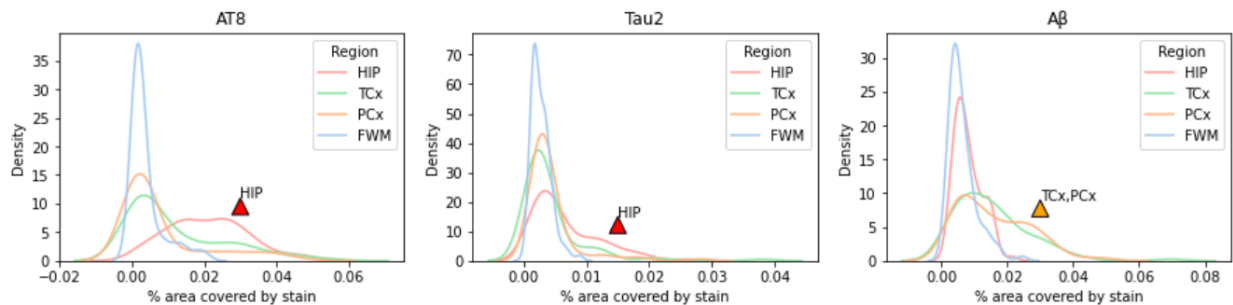


Figure 12

In Figure 13, we used AT8 IHC level to represent pTau protein level in the brain tissues. AT8 IHC and pTau are strongly correlated (Pearson correlation coefficient is 0.68). We scaled the AT8 IHC reading to values between 0 and 1, then evenly divide the range to 5 bins. So, the Bin 1 represents the range between 0.0 and 0.20. Bin 5 represents the range between 0.81 and 1.0. We can observe that later Braak stages are associated with higher pTau levels. 2/3 of the samples with Braak Stage 6 have pTau protein level 2 and above while 1/6 of the samples with Braak Stage 1 have pTau protein level 2 and above. We also noticed that more samples with higher pTau protein level have dementia. 38% of the samples with pTau protein level 1 have dementia while 86% of the samples with pTau level 5 have dementia. It shows moderate correlation between pTau level and dementia status.

## Braak stage, pTau protein and Dementia Status

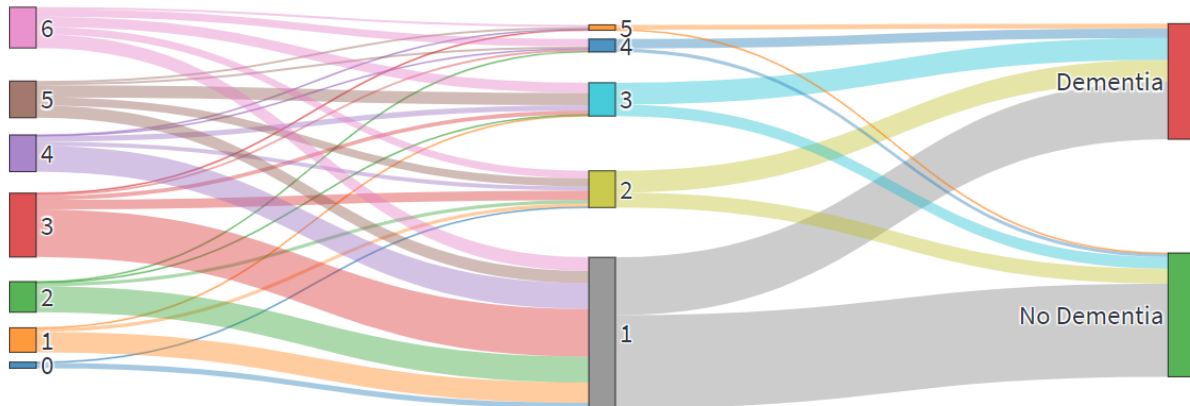


Figure 13

### 4.2 Machine Learning

This section talks about machine learning model training, validation and predictions making.

One of the important aspects of this project is that we will be able to predict probability that a person might struggle with dementia as they grow old. This prediction is based on the tests that can be performed on live patients.

### Tests that can be performed on live patients:

1. **Histology and immunohistochemistry (IHC):** Image data and quantitative image metrics to assess  $\beta$ -amyloid, tau, and  $\alpha$ -synuclein pathologies as well as the overall local pathological state of tissue samples from each donor.

a.  **$\beta$ -amyloid** – Membrane Protein imp role in neural growth & repair, plaques found in Alzheimer's

b. **tau** – stabilize internal skeleton (tube for nutrients) of Neuron, detach from tubules blocks neuron transport system

c.  **$\alpha$ -synuclein** – regulate neurotransmitter release, forms lewy bodies (abnormal shape protein) with SNCA gene mutation

How is amyloid plaque diagnosed in live patients?

The **blood test** looked for two forms of beta-amyloid protein: beta-amyloid 42 and beta-amyloid 40. When beta-amyloid begins to build up, the ratio between the two proteins changes, and the blood test detects this.

Bone marrow tests or other small biopsy samples of tissue or organs can positively confirm the diagnosis of amyloidosis

Live patient testing for tau protein - **PET scans of the brain and lab tests of spinal fluid for tau protein**

## 2. What is Apo E4?

The apolipoprotein E4 (ApoE) allele is **the strongest known genetic risk factor for sporadic AD** (Corder et al., 1993). ApoE is primarily produced by astrocytes, however, its role in pathology remains unclear given that it appears to be involved in both aggregation and clearance

The APOE protein **helps carry cholesterol and other types of fat in the bloodstream.**

Source: <https://www.alzheimersorganization.org/alzheimers-test>

This effect is additive in that one copy of e4 (e2/e4 or e3/e4) carries some increased risk and two copies of e4 (e4/e4) are associated with even more of a risk of developing AD. It is important to note, however, that we are talking about the risk relative to other people at the same age with fewer copies of e4. In terms of lifetime risk, most individuals with *APOE* e4 will never develop AD and there are many people with AD who are e4 negative

## 3. What is CERAD Test?

The Consortium to Establish a Registry for Alzheimer's Disease neuropsychological battery (CERAD-NB) is **a relatively brief standardized test battery designed to measure primary cognitive deficits in AD**

**(clocks are scored on a 4-point scale: normal, mild, moderate, severe impairment, with examples provided for each level),**

NIA-Reagan diagnosis of Alzheimer's disease is **based on consensus recommendations for postmortem diagnosis of Alzheimer's disease.** The criteria rely on both neurofibrillary tangles (Braak) and neuritic plaques (CERAD).

## 4. **Braak staging** refers to two methods used to classify the degree of pathology in Parkinson's disease and Alzheimer's disease

**This is obtained by performing an autopsy of the brain**

Braak staging in AD has six stages — I through VI. The staging focuses on the location of NFTs. Stages I and II are when the NFTs are limited to the transentorhinal region of the brain. Stages III and IV are when the NFTs are in the limbic regions, which includes the hippocampus.

Braak stages I and II are used when neurofibrillary tangle involvement is confined mainly to the transentorhinal region of the brain, stages III and IV when there is also involvement of limbic regions such as the hippocampus, and V and VI when there is **extensive neocortical involvement**.

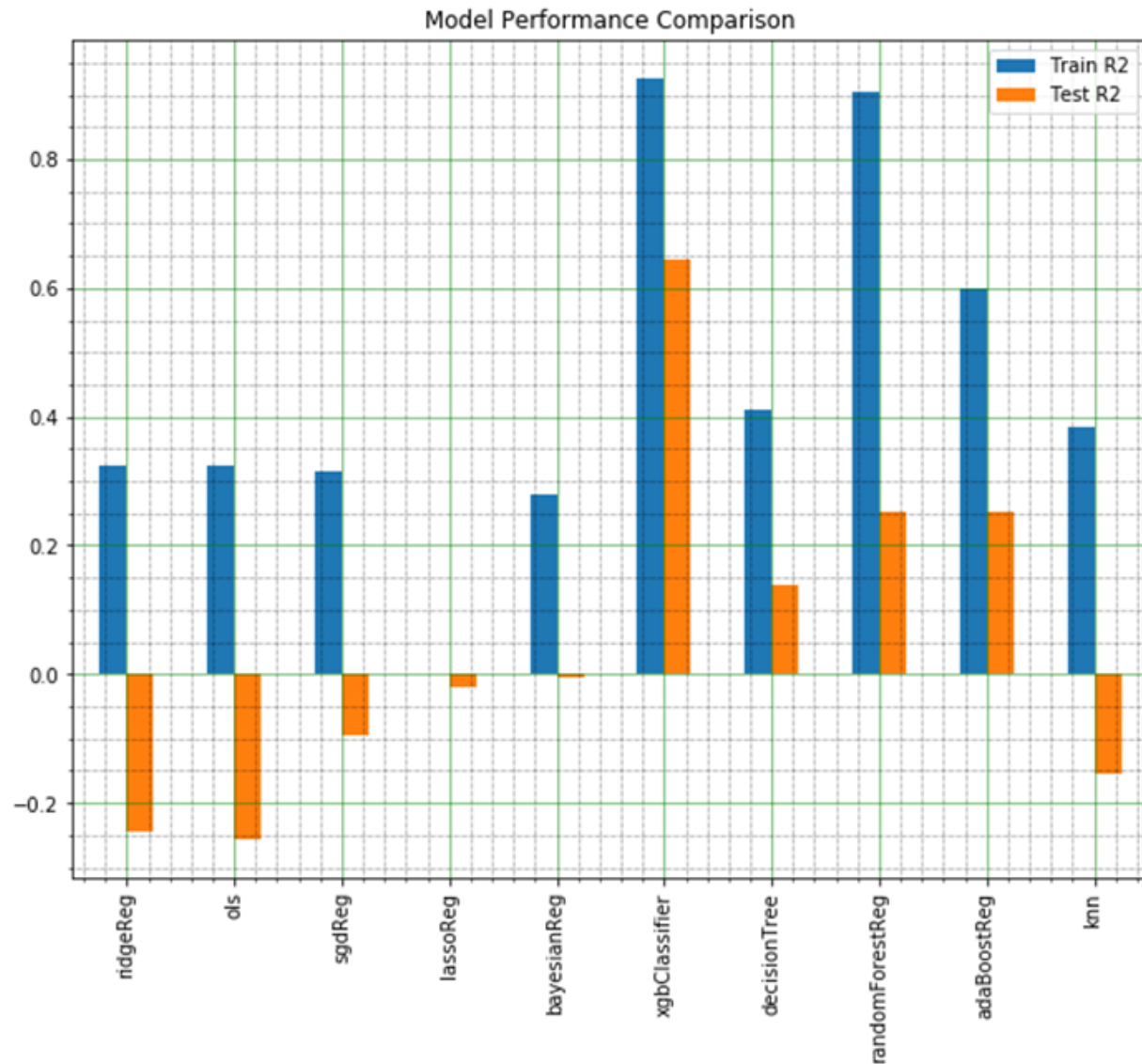
## Selecting a prediction model

We started by combining all the different files into one to bring all the parameters against each donor. After we did the required EDA (discussed above), we drilled down to 34 features for the model. On very high level, we proceeded by splitting the entire data into test and train (20%/ 80%), perform normalization, compare accuracy of different models to final select one, perform feature importance and finally do a 5-fold cross validation to come up with final model.

```
In [16]:  X_train.shape, y_train.shape, X_test.shape, y_test.shape
Out[16]: ((301, 34), (301,)), (76, 34), (76,))
```

### But how did we come up to final model?

This was a one of the big challenges that we faced. Generally, for any machine learning, models are trained on a good number of data records but this study being a cohort study, it had a total of 377 donors. Training a model that can work great with such a small sample is critical for good prediction. Hence, we started by experimenting with many ML models and compare their accuracy score. From our secondary research, we came up with 10 different models and tested them on this data with default parameters settings to get a general sense of what type of model can work and which model might not be a good choice. From these models, we chose top 3 models for parameter tuning. As it is visible from below, **top models that we moved ahead with is XGBClassifier, adaBoost Regression and Random forest regression**. Post parameter tuning we again compared the r-squared values of these 3 models and we went ahead with top performing model which is XGBClassifier and using its predict\_prob for final predictions.



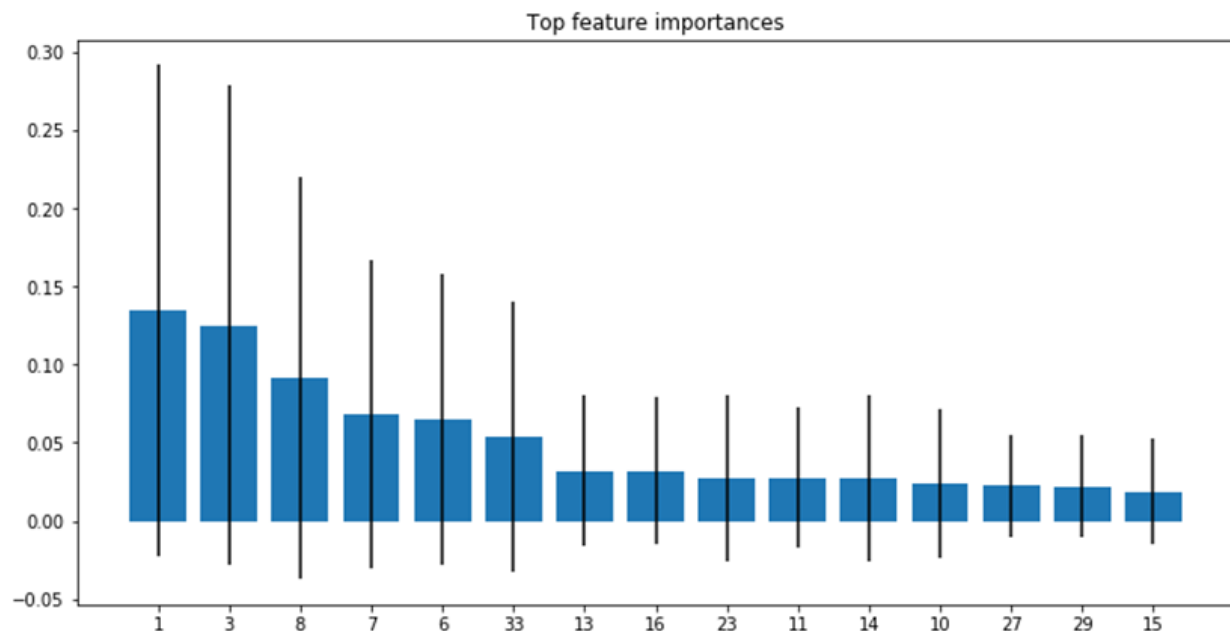
Once it was evident that our **final model will be XGBClassifier**, we worked on its feature importance, cross-validation along with parameter tuning that we initially did.

### What is XGBoost?

The XGBoost stands for eXtreme Gradient Boosting, which is a boosting algorithm based on gradient boosted decision trees algorithm. Another challenge that comes with modeling on small dataset is overfitting. XGBoost applies a better regularization technique to reduce overfitting, and it is one of the differences from the gradient boosting. This model falls in the category of ensembles of decision trees.

### Feature Importance

A benefit of using ensembles of decision tree methods like gradient boosting is that they can automatically provide estimates of feature importance from a trained predictive model. In addition of using feature importance from model, we also tested feature importance with help of ExtraTreesClassifier provided in sklearn.ensemble which is used to determine feature importance especially for decision trees. Out of initial 34 features, we selected top 15 features which explained our model the best



### Cross-validation

We performed a 5-fold cross validation with these top 15 features on our final model of XGBClassifier. **We are able to reach a train r-square value of 95% and test r-square value of 70%.** Which is a pretty good model on a such a small dataset. From this here we created a pickle file of trained model as well as normalizer that is now deployed on the website. A user can enter their test results and at backend we will perform same normalizer that is earlier trained and then predict\_probability, again from the earlier trained model.

### 4.3 Website

We created a website for our study. (<https://project-dementia.herokuapp.com/>) It includes many nice features. It allows users to explore the dataset used in this study (See Figure 14). You can select different data files, choose different pairs of variables for the scatterplots.

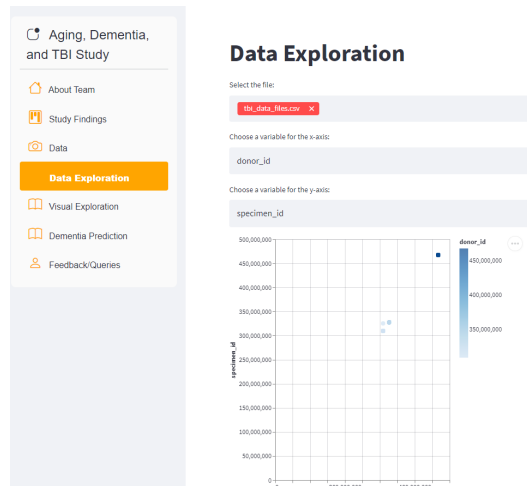


Figure 14

The website also shows some visualizations to help users understand the relationship between variables. It includes parallel coordinates plots (Figure 15) and Sankey diagrams (Figure 16).

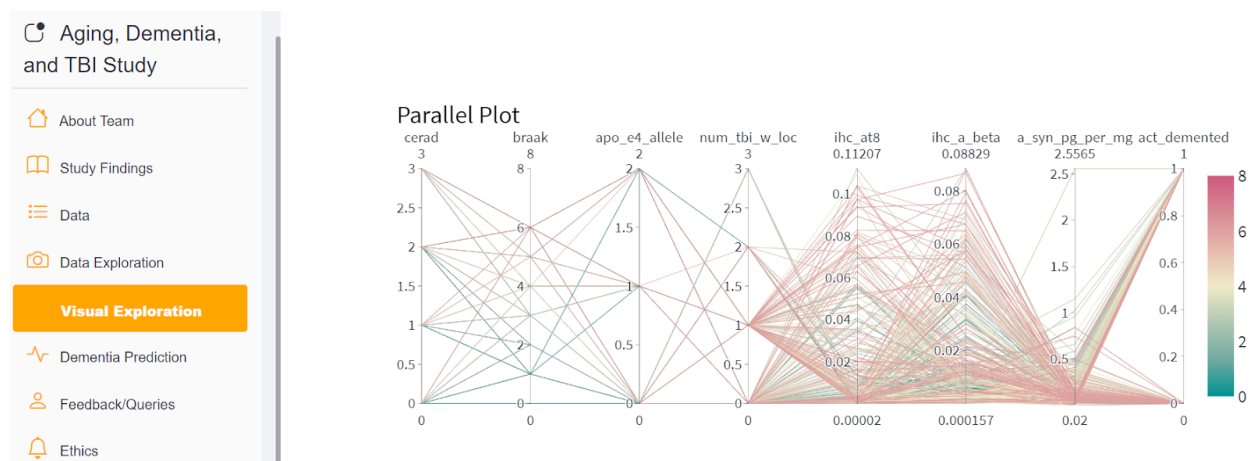


Figure 15

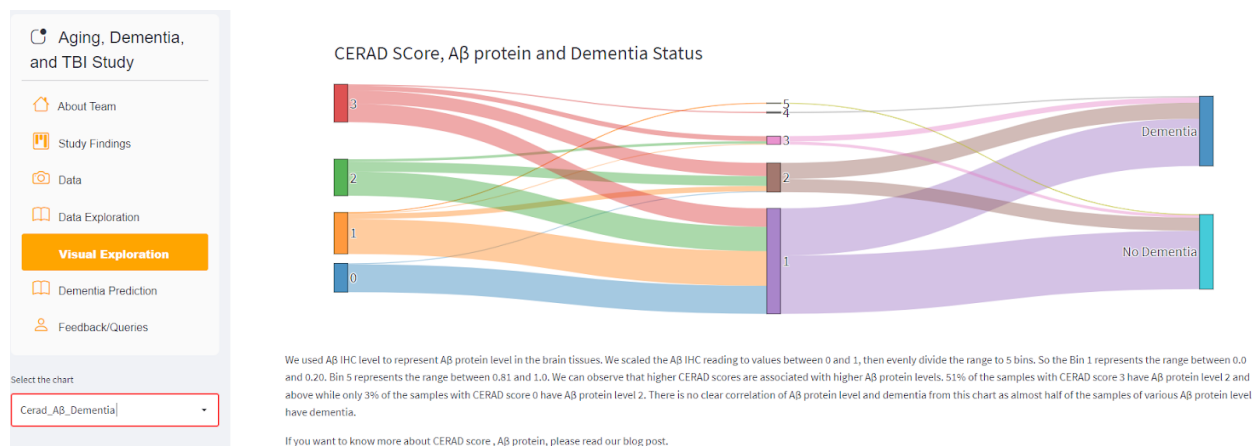




Figure 16

Most fascinating part of the website is the predictive model. Users can upload a CSV file or enter the metrics of the test results and the predictive model provides user a probability of getting dementia based on the input (Figure 17).

The screenshot shows a web application titled "Brain Dementia Prediction Model". On the left is a sidebar menu with the following items: "Aging, Dementia, and TBI Study" (active), "About Team", "Study Findings", "Data", "Data Exploration", "Visual Exploration", "Dementia Prediction" (highlighted in orange), and "Feedback/Queries". The main content area has a header "Brain Dementia Prediction Model" in orange. Below it is an "Upload File(csv)" section with a "Drag and drop file here" area (Limit 200MB per file • CSV) and a "Browse files" button. The "Enter Values:" section contains a grid of input fields for the following variables: cerad, braak, ihc\_tau2\_fpe, ihc\_at8\_fpe, ihc\_at8, ihc\_a\_beta\_fpe, ihc\_a\_beta, ihc\_gfap\_fpe, ptau\_ng\_per\_mg, vegf\_pg\_per\_mg, ptau\_over\_tau\_ratio, mcp\_1\_pg\_per\_mg, ab42\_over\_ab42\_ratio, a\_syn\_pg\_per\_mg, and ab42\_pg\_per\_mg. A "Submit" button is located at the bottom left of the input grid.

Figure 17

#### 4.4 Code Repository

- Data analysis and model training code: <https://github.com/xu4/mads-capstone>
- Website code: [https://github.com/singhsud2157/project\\_capstone](https://github.com/singhsud2157/project_capstone)

#### 5. Ethics

As we are dealing with medical data, we need to think of ethical implications. Patient data have conventionally been thought to be well protected by the privacy laws outlined in the United States. But there are great concerns that shared patient data or data voluntarily provided by patients for research may be exploited for commercial interests. We as researchers should consider participants' privacy and security while using shared data through new technologies such as artificial intelligence and other remote technologies.

#### 6. Conclusion/Discussion

In this study, we analyzed the relationships of pTau protein and A $\beta$  protein levels with Braak stage, CERAD score, number of TBI in four brain regions from 107 well-characterized donors from the ACT cohort.

We confirm known associations between pTau and A $\beta$  pathologies and dementia. pTau and A $\beta$  have moderate correlations in hippocampus and cortex while having no or very weak correlation in frontal white matters. The correlations of the pTau and A $\beta$  proteins in later Braak stages (5 and 6) are much stronger than early stages. We did not find correlation between the number of TBI and dementia status.

## **7. Future work**

The relationship between tau pathology and dementia status is not consistent in different age group and we would like to do further research on that relationship in the subsets of the cohort.

The ACT study dataset also contains RNA sequencing data from the 107 donors. In the future we would like to study the relationship of gene expression and dementia using the RNA sequencing data. We may identify genes significantly associated with dementia and rank them. We could also study the relationship of gene expression with pTau and A $\beta$  pathology and inflammation.

## **8. References**

<http://aging.brain-map.org/overview/home>

<https://github.com/AllenInstitute/agedbrain>

<https://elifesciences.org/articles/31126#fig4sdata1>