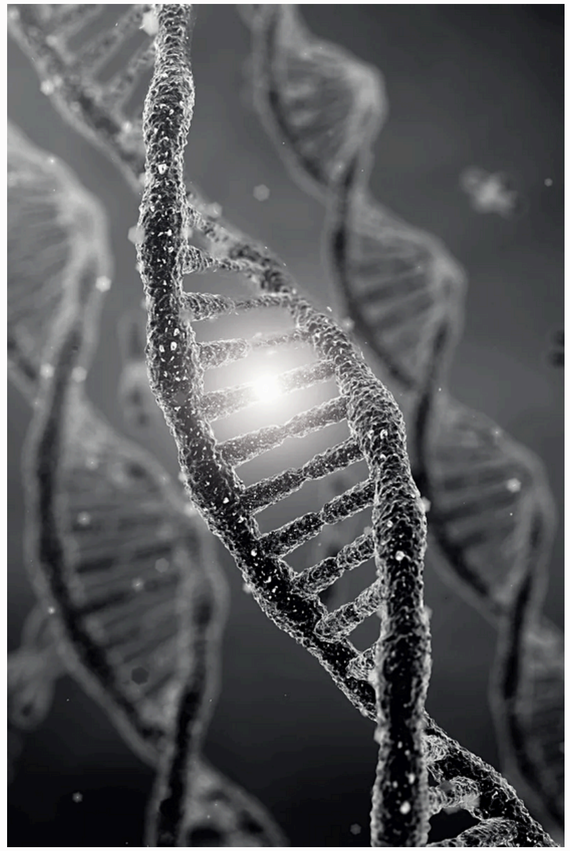


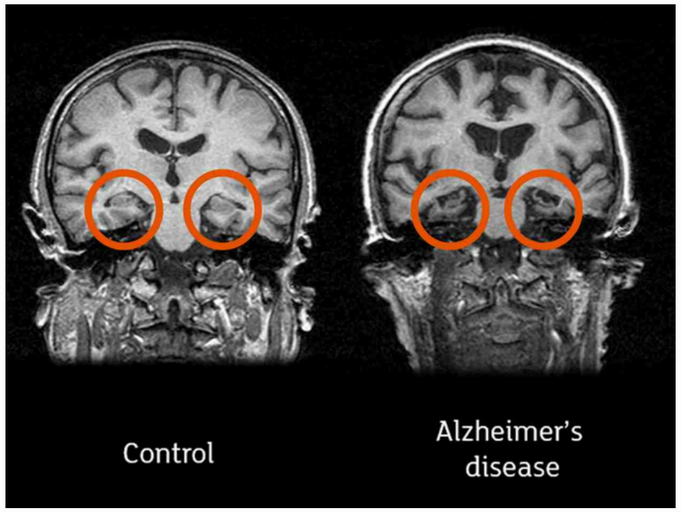
MADS Seminar Series

Predicting Biological Age in Alzheimer's Disease Using Machine Learning

John Seibert (CU-Bloomsburg)

Joint research with:
Harley Nguyen (UC Irvine undergrad student)
Wei Vivian Li (UC Riverside statistics faculty mentor)
Wenxiu Ma (UC Riverside statistics faculty mentor)
Xinping Cui (UC Riverside statistics faculty mentor)

September 23, 2025



Agenda

1. Intro
2. Background
3. Scientific question and hypothesis
4. Methodology
5. Results
6. Discussion
7. Q&A

The program

- **About**
  - REU: Research Experience for Undergraduates
  - Nationwide programs sponsored by the National Science Foundation
- **UC Riverside's program in particular**
  - 8-week research experience in data science (focused on "genomic data mining")
  - Presented work at SoCal REU Symposium @ Harvey Mudd College and at UCR Data Science High School Summer Camp
  - Received professional development, including a graduate school application workshop and industry advice from a panel of data scientists
  - All else, 8 weeks of not worrying about food and housing, plus an excuse to travel!

My focus

- **Point of clarity**
  - “Genomic data mining” was statistical machine learning using transcriptomic data.
- **Terminology I use a lot here:**
  - **Machine learning:** models learn and improve from data without being explicitly programmed
    - Explain the textbook email spam example
    - Statistical theory deeply rooted in ML (regression, decision trees, etc.)
  - **Thus, you can argue statistical machine learning**: statistical methods and theory as the foundational mathematical framework for building and validating machine learning model
  - **Transcriptomic data:** complete set of RNA transcripts present in a cell, tissue, or organism at a specific moment

On my work

- **What is Alzheimer's Disease (AD)?**
  - Alzheimer's disease is a growing health challenge. (Zhang, 2021)
  - characterized by a gradual decline in cognitive function and structural changes throughout the brain (Wong, 2020; Zhang, 2021)
  - Impacts cognitive centers, particularly the frontal cortex
- **What is transcriptomics, and why utilize this approach?**
  - Transcriptomics examines age-related changes in gene expression
  - Why use this technique? Aging is an asynchronous process across tissues, with distinct periods of major transcriptional changes (Schneider, 2024)
  - Brain-age delta (difference in chronological age and predicted biological age) is a clinically relevant marker of brain aging, associating with AD pathology even in non-demented individuals (Cumplido, 2023)

Background

- **Alzheimer’s Disease (AD)**
  - Alzheimer’s Disease (AD) causes severe memory and cognitive decline.
  - Aging is a major risk factor. AD may involve accelerated molecular aging, measurable through transcriptomic data.
  - Affects 4.5M Americans (2000) → Projected 13.2M by 2050
  - Estimated $1+ trillion in treatment costs

- **Biological age ≠ Chronological age:** Biological age reflects molecular state of aging estimated from age-responsive gene expression rather than years lived.
  - Biological age can be inferred from gene expression (transcriptomic age).

Scientific Question

Main question:

Do transcriptomic profiles reveal accelerated aging in Alzheimer's Disease brains compared to healthy controls?

Sub-questions:

1. *Can we predict biological age from gene expression data in healthy individuals?*
2. *Do AD brains show higher predicted transcriptomic age than chronological age?*

Hypothesis: AD patients show accelerated transcriptomic aging.

Literature Review

voyAGEr Study (2024)

- Explore how gene expression changes with age in specific tissues and between sexes.
- Dive deep into individual genes to see their aging patterns, or zoom out to understand how entire tissues are affected.
- Understand groups of genes that work together and how they relate to biological processes and diseases.

Horvath epigenetic clock study (2013)

- Created a universal "epigenetic clock" that can accurately estimate the age of most human and chimpanzee tissues
- Tested his theory on 8,000 non-cancer samples from 82 Illumina DNA methylation array datasets, covering 51 healthy tissues and cell types
- Found that cancer tissues show a significant "age acceleration," making them epigenetically much older than they actually are (36 years older, in fact!)

Datasets

I. GTEx v8

Tissue: Brain (Frontal Cortex, BA9)

Samples: Over **900** healthy donors

Platform: Bulk RNA-seq

Source: GTEx Portal

- Sample metadata: GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt
- Subject phenotypes: GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt
- Gene counts: Gene Read Count by Tissue under GTEx Analysis V8

II. GSE125583

Samples: **289** individuals (**~70** controls, **~219** Alzheimer's disease cases)

Platform: Bulk RNA-seq

Source: GEO - GSE125583

Gene Selection Process (voyAGEr Preprocessing)

Tissue Filtering	Age Filtering	Gene Identification	Gene Selection
Filtered to select only “Brain - Frontal Cortex (BA9)” tissue	Narrowed age range of GTEx data from 20-70	Selected genes with just significant age effect but no sex effect, based on Benjamini-Hochberg procedure	Selections top genes sorted by adjusted p-value

Methodology

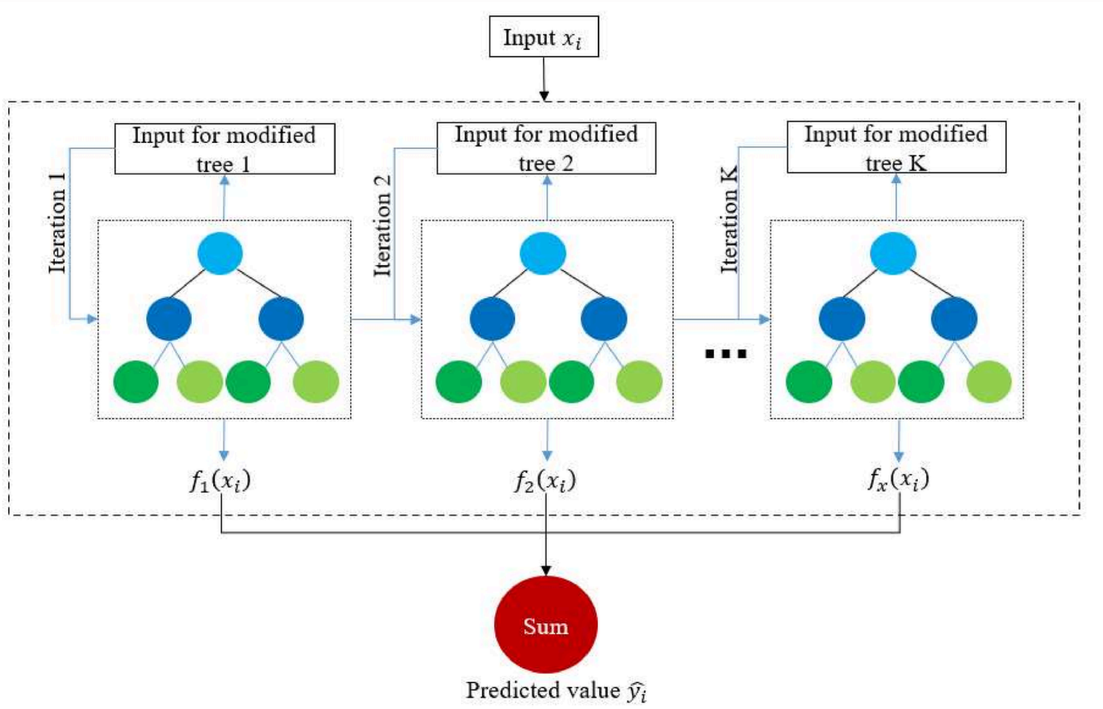
- Identified age-responsive genes sorted by adjusted p-value (controlled for false discovery rate) from GTEx v8 using voyAGEr-style filtering and linear modeling (age + sex).
- Downloaded and filtered normalized expression and metadata from GSE125583.
- Samples were split into control (for training) and AD (for testing).
- Expression data were log2(TPM + 1) transformed and z-score standardized.
- Applied trained models to AD samples to estimate biological age. Age acceleration was calculated as: **Age Acceleration = Predicted Age - Chronological Age**
- Compared age acceleration in AD vs. controls using t-test

Top 100 Genes

GFAP	1.1366870	0.050513	0.206998	0.000279	0.594854	1.8209726	0.069844	Age_Only
GPR26	8.0872392	-0.05869	0.16558	0.021155	-0.46051	6.4778786	0.304488	Age_Only
STUM	2.2666875	-0.03933	0.147533	0.116428	-0.2166	0.000121	0.476975	Age_Only
LINC0147(	1.9722401	-0.05734	0.13214	0.069033	-0.39559	0.00079	0.418981	Age_Only
NREP	5.5078573	-0.03097	0.120233	0.119309	-0.19045	0.001471	0.478609	Age_Only
DOK6	5.0314914	-0.0402	0.125537	0.055017	-0.30383	0.001471	0.392596	Age_Only
FOXJ1	6.5851773	0.049373	0.115689	0.208881	0.246473	0.001507	0.559016	Age_Only
OTOF	1.5984710	-0.03024	0.103171	0.729027	-0.04314	0.002397	0.892524	Age_Only
CTD-2015I	1.6403083	-0.02695	0.110849	0.118327	-0.17405	0.002397	0.477646	Age_Only
C8orf34	1.6462190	-0.04208	0.107517	0.21617	-0.21501	0.002397	0.564933	Age_Only
PTHLH	1.2542447	-0.03948	0.129812	0.008625	-0.42629	0.002397	0.230943	Age_Only
CALB1	1.9397056	-0.05194	0.125659	0.009339	-0.56567	0.00259	0.237489	Age_Only
OTOS	2.1642336	0.049174	0.106522	0.166227	0.284879	0.002667	0.52441	Age_Only
KIF21B	3.9012749	-0.02762	0.114379	0.020848	-0.27571	0.002878	0.304488	Age_Only
CYP26B1	4.6391588	-0.04325	0.094951	0.457981	-0.13893	0.002878	0.74354	Age_Only
AC141928	3.7150472	-0.02713	0.101975	0.160954	-0.16322	0.002878	0.519646	Age_Only
CAMK4	4.0064391	-0.03662	0.112859	0.0253	-0.35408	0.002878	0.322615	Age_Only
NEURL1B	4.2403296	-0.02286	0.099869	0.191067	-0.12902	0.002878	0.54329	Age_Only
GUCA1A	3.8875495	-0.02884	0.095279	0.711601	-0.04574	0.002878	0.883093	Age_Only
HEBP2	4.5802634	0.021954	0.124	0.004252	0.274219	0.002878	0.180316	Age_Only
GRIN3A	3.8711047	-0.0464	0.113542	0.023858	-0.45256	0.002878	0.312702	Age_Only
KAZALD1	4.6171236	-0.02804	0.124351	0.004007	-0.35274	0.002878	0.179063	Age_Only
PSTPIP1	4.6707950	0.021438	0.155745	5.0135954	0.384288	0.002878	0.022795	Age_Only
MMD	4.5771377	-0.02336	0.100523	0.150619	-0.14552	0.002878	0.509377	Age_Only
DLGAP1-A	3.8112491	-0.03237	0.114268	0.02183	-0.32032	0.002878	0.306588	Age_Only
AC000095	3.4001693	-0.02992	0.13276	0.001734	-0.40456	0.002878	0.136838	Age_Only
COL25A1	5.1663925	-0.04167	0.101427	0.107707	-0.29241	0.003065	0.467774	Age_Only
PPP4R4	6.3318230	-0.03303	0.117084	0.007776	-0.38966	0.003623	0.222288	Age_Only

XGBoost

- Incorporates regularization, handles missing values, and cross-validates at each iteration (Chen & Guestrin, 2016)
- Enables early stopping, finding the optimal number of iterations (Chen & Guestrin, 2016)



Elastic Net

- Elastic Net blends Lasso (L1) and Ridge (L2) regularization to handle correlated gene expression data (Zou & Hastie, 2005).
- Performs variable selection and shrinks coefficients to prevent overfitting, ideal for "high p, low n" scenarios.
- Used in Horvath’s epigenetic clock to select age-informative CpGs across tissues (Horvath, 2013).
  - Applied to transcriptomic clocks and Alzheimer’s classification:
  - Abdullah et al. (2022) used EN logistic regression on blood mRNA, achieving 81.6% accuracy in distinguishing AD from controls.
  - EN handles multicollinearity and outputs interpretable gene panels, aiding biological insight.

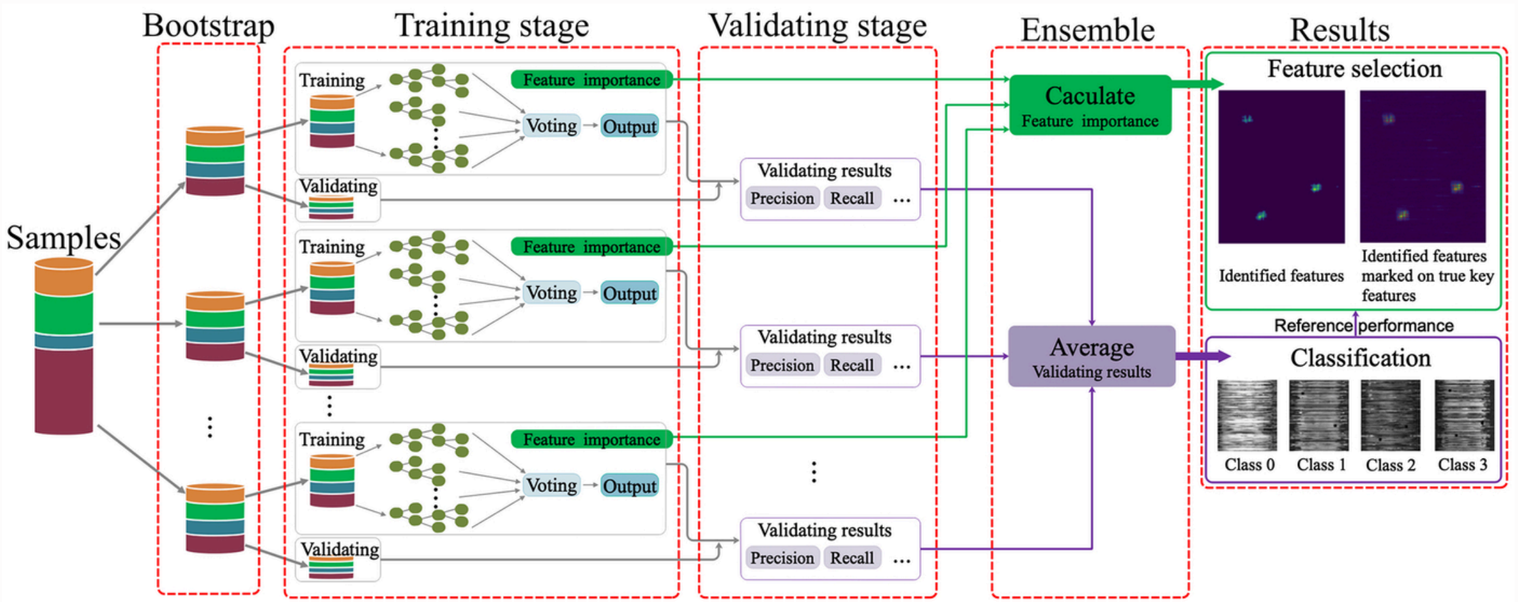
$$\text{Elastic Net Loss} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2\right)$$

Here:

1. $\frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: Mean squared error (MSE), the regression loss term measuring the difference between predicted and actual values.
2. λ: Regularization parameter, controlling the overall strength of the regularization.
3. α: Mixing parameter, determining the balance between L1 (Lasso) and L2 (Ridge) penalties.
4. $\sum_{j=1}^{p}|\beta_j|$: L1 penalty, inducing sparsity.
5. $\sum_{j=1}^{p}\beta_j^2$: L2 penalty, encouraging small coefficients.

Random Forest

- Handles high-dimensional, noisy gene expression data effectively.
- Captures complex gene-gene interactions and nonlinear effects.
- Used in Alzheimer's transcriptomic classification tasks:
  - Wu et al. (2021) applied Integrated Multiple RFs to distinguish AD from controls with transcriptomic data.
  - There is “AITeQ” – an RF-based tool achieving high AD classification using only 5 genes (Ahammad et al., 2023).



Model Training Process

Data Collection

- Dataset: GSE125583
- Tissue: Brain Frontal Cortex (BA9)
- Groups: AD vs Control
- Age range: 20-80 years

Gene Filtering

- Starting genes: 20,000+
- Selection criteria: $R^2 > 0.1$, $p < 0.05$
- Final selection: 400 genes
- Top genes: GFAP, GPR26, STUM, PTHLH

Preprocessing

- Transformation: Log2(TPM + 1)
- Sex encoding: Male=0, Female=1
- Quality control: Remove outliers
- Missing data: Imputation or removal

Model Training Process (cont.)

Predictor Matrix Creation

- Reference: Control samples only
- Method: Z-score (mean=0, SD=1)
- Purpose: Prevent feature dominance
- Validation: Check distribution

Model Training

- Cross-validation: 5-fold
- Models: Random Forest, XGBoost, Elastic Net
- Metrics: R^2, RMSE, MAE
- Validation: Prevent overfitting

Comprehensive Analysis

- Age prediction: All samples
- Age acceleration: Predicted - Actual
- Statistical tests: T-tests, Wilcoxon
- Visualization: Boxplots, scatterplots

Model Performance

XGBoost (5-fold CV) Performance:

R-squared (R^2): 0.283

Root Mean Squared Error (RMSE): 5.797

Mean Absolute Error (MAE): 4.569

Random Forest (5-fold CV) Performance:

R-squared (R^2): 0.039

Root Mean Squared Error (RMSE): 6.451

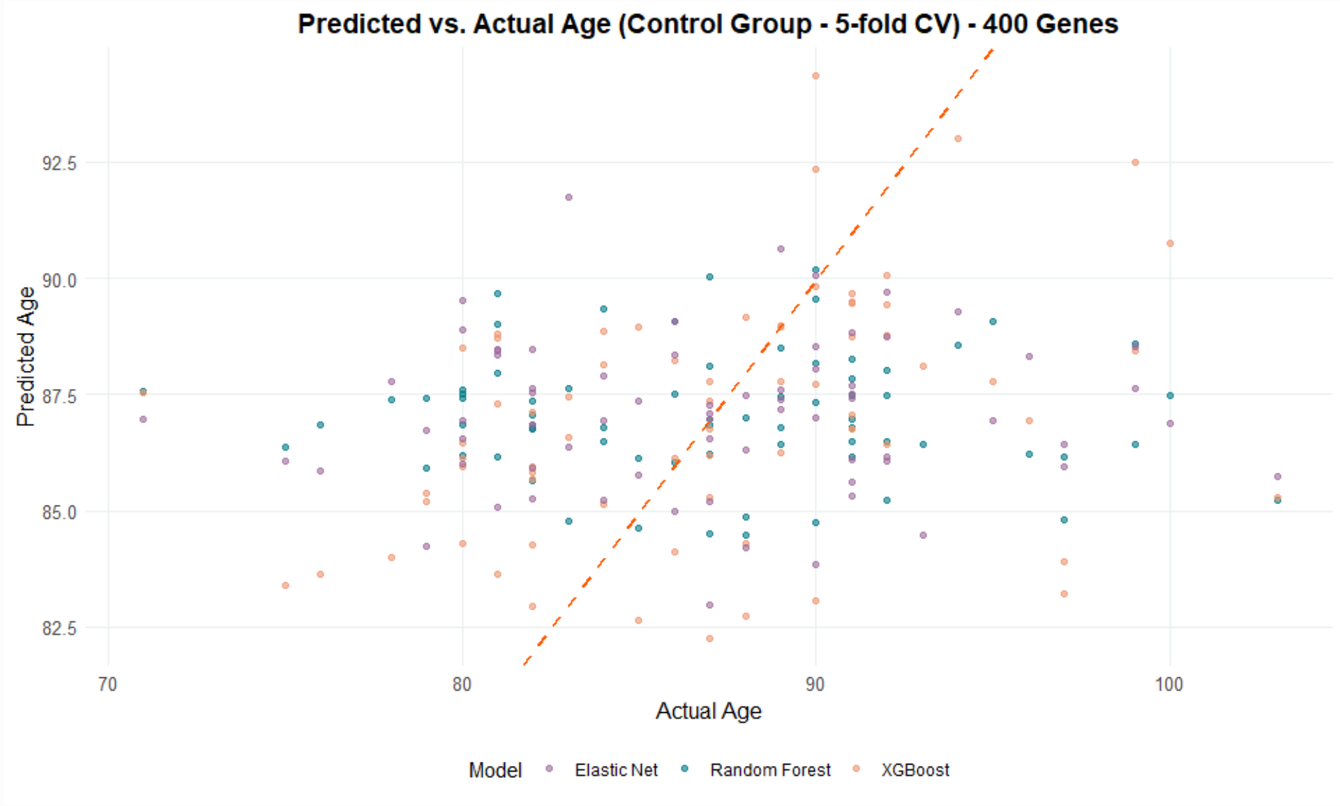
Mean Absolute Error (MAE): 5.197

Elastic Net (5-fold CV) Performance:

R-squared (R^2): 0.053

Root Mean Squared Error (RMSE): 6.41

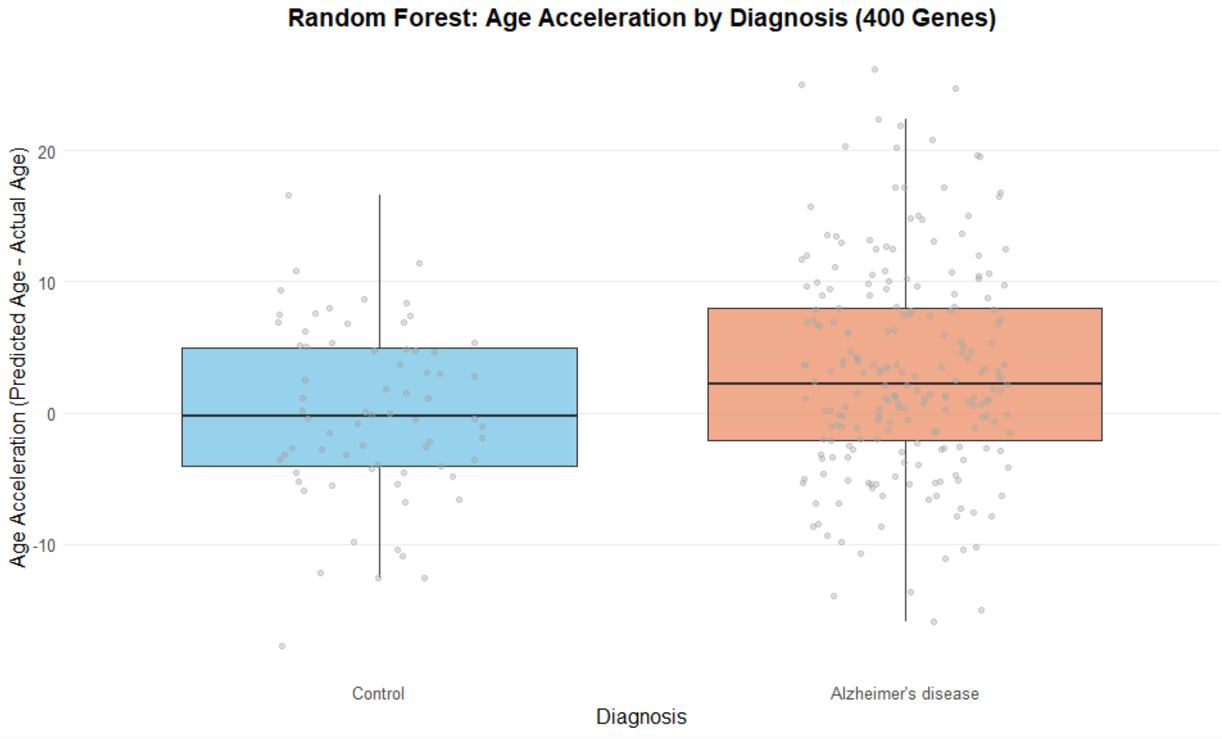
Mean Absolute Error (MAE): 5.198



Random Forest

Random Forest Age Acceleration Statistics (Control):

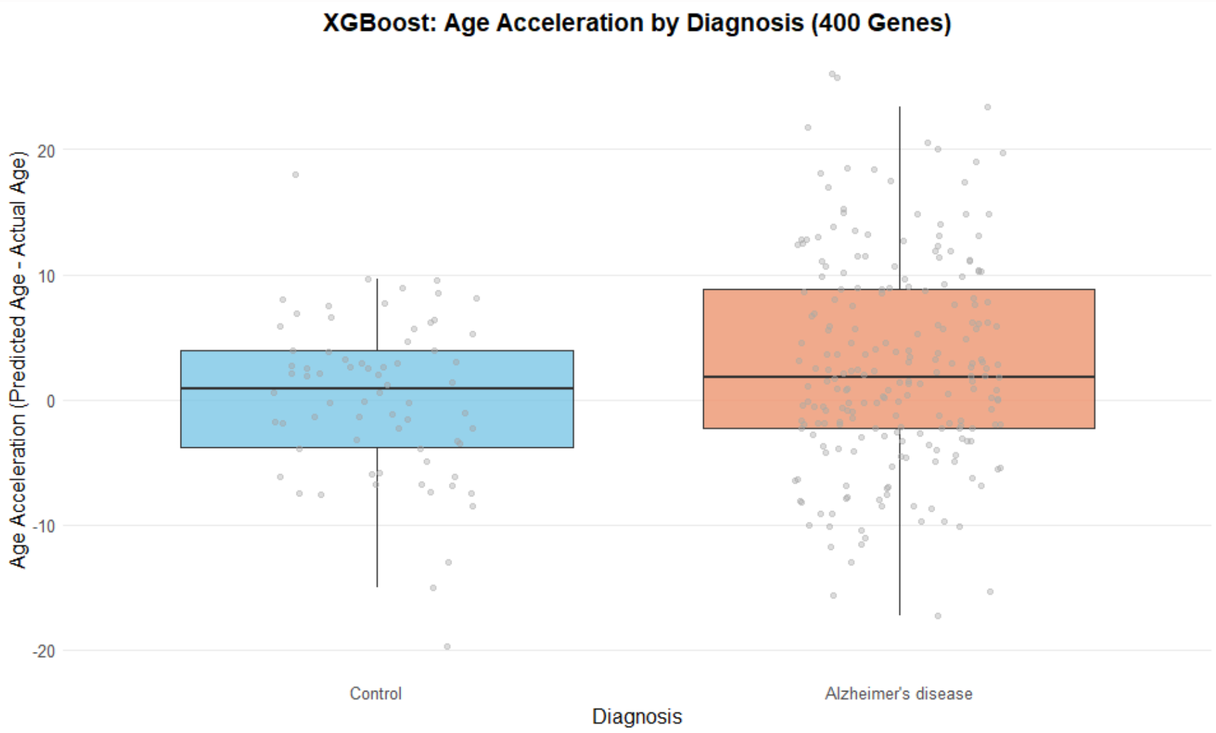
- Mean: 0.045
- Median: -0.265
- SD: 6.523



XG Boost

XGBoost Age Acceleration Summary (Control):

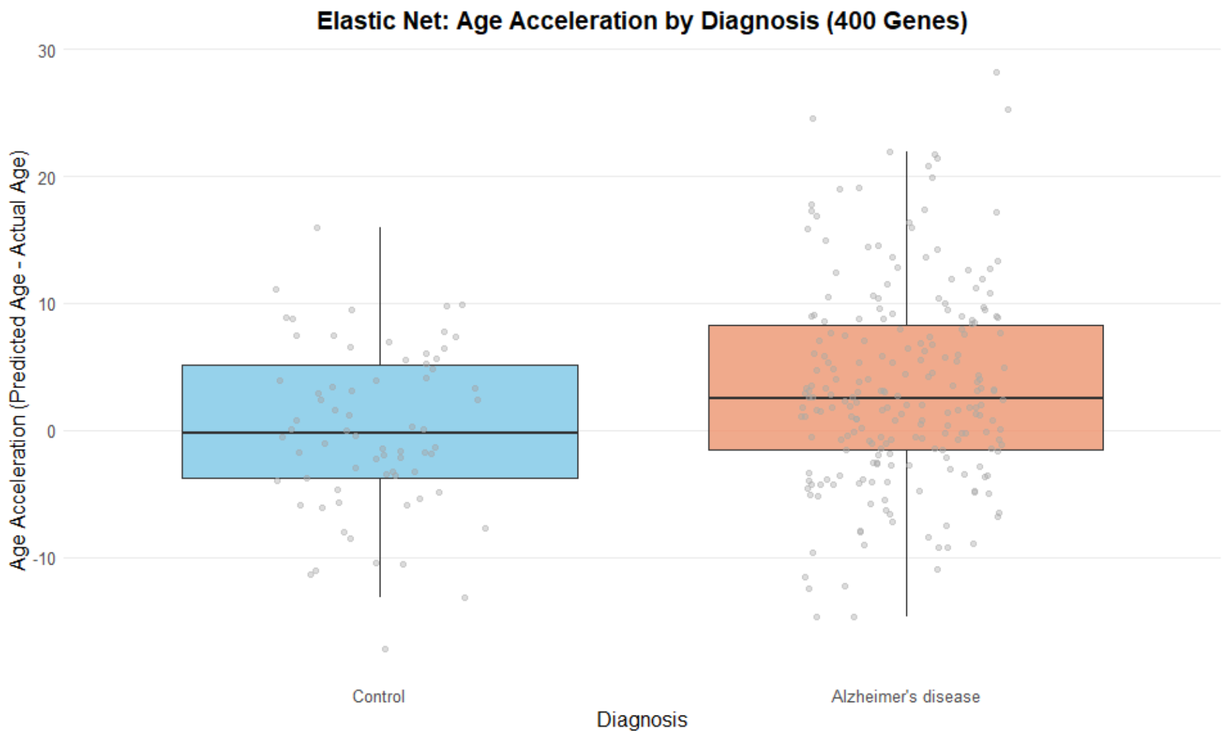
- Mean: 0.01
- Median: -0.036
- SD: 5.917



Elastic Net

Elastic Net Age Acceleration Summary (Control):

- Mean: 0.072
- Median: -0.234
- SD: 6.482



Non-parametric tests

Wilcoxon Signed Rank Test

- **Random Forest:**
  - V = 1283, p-value = 0.8149
  - alternative hypothesis: true location is not equal to 0
- **XG Boost:**
  - V = 1344, p-value = 0.5545
  - alternative hypothesis: true location is not equal to 0
- **Elastic Net:**
  - V = 1277, p-value = 0.8423
  - alternative hypothesis: true location is not equal to 0

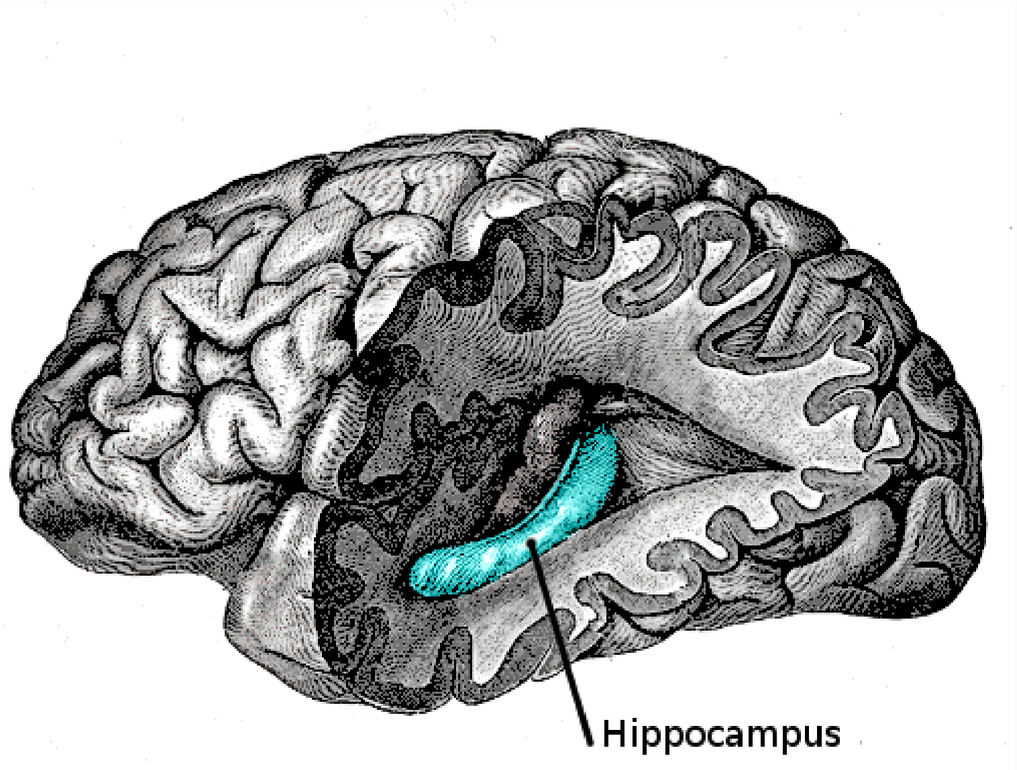
Wilcoxon Rank Sum Test

- **Random Forest:**
  - W = 9253, p-value = 0.003721
  - alternative hypothesis: true location shift is greater than 0
- **XG Boost:**
  - W = 8782, p-value = 0.02875
  - alternative hypothesis: true location shift is greater than 0
- **Elastic Net:**
  - W = 9336, p-value = 0.002452
  - alternative hypothesis: true location shift is greater than 0

New Approach: Hippocampus Age-Responsive Gene Analysis

Research Objective

- **Goal:** Identify age-responsive genes in hippocampus tissue from GTEx v10 dataset for biological age prediction and train the predicting modeling using samples from GSE 173955
- **Challenge:** Develop an efficient & strong pipeline for gene selection and validation
- **Approach:** Obtain better result for predicting modeling



Experimental Reflection

The Core Question

Accelerated aging does show in AD. Ultimately, we need to explore many more tissues, like how Horvath used 51 in his seminal paper on DNA methylation.

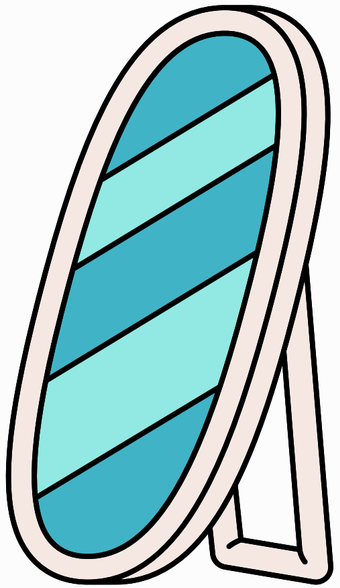
Gene Expansion

We showed in earlier presentations the results of models with less genes, with higher R^2. Initially, we thought more genes would lower R^2 but this is not the case.

Pipelines

The new approach is our third pipeline. We are still in progress for future work. We have learned that the key of research is keeping asking questions.

Experience Reflection



Always Ask Questions

Being Persistent and Conquer The Challenge

Embrace the Learning Curve

Collaboration is Key

Future Direction

- **Expanding Sample Dataset:** Discover other open access GSE dataset that have larger samples.
- **Multi-tissue Analysis:** Extend this approach to other brain regions and peripheral tissues
- **Longitudinal Studies:** Track age acceleration over time to assess disease progression
- **Therapeutic Targets:** Investigate the identified genes as potential targets for AD intervention

References

This research was supported by the National Science Foundation under REU Site grant no. OAC 2244480.

THANK YOU EVERYONE FOR LISTENING TODAY!

I would like to send my gratitude to Professors Wei Vivian Li, Wenxiu Ma, Xinping Cui for guiding us through what has been meaningful project over my undergraduate career. I am also thankful for the contributions Harley Nguyen gave to this work. Beyond that, I am very grateful for all involved in the REU program, especially Drs. Jia Chen, Joyce Fu, and Vagelis Papalexakis, for running it and having any logistics go smoothly.