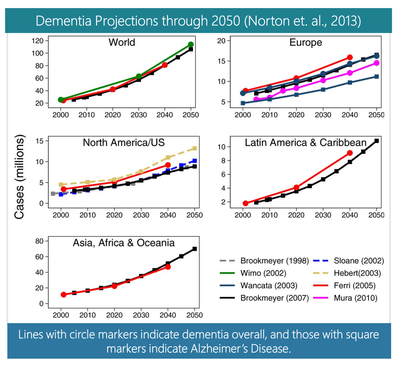
A Machine Learning Approach to Identifying Blood-Based Biomarkers for Differential Diagnosis of Alzheimer's Disease

Anjali Sreenivas | Nikola Tesla STEM High School, Redmond, WA | CBIO038

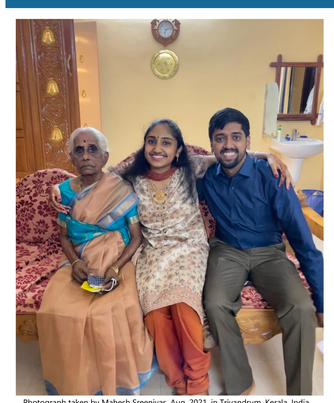
Introduction

Background

- Alzheimer's Disease (AD) is a devastating age-related neurodegenerative disorder of gradual onset
  - Clinically characterized by cognitive deterioration and memory loss
  - Central pathological hallmarks: amyloid-β senile plaques & neurofibrillary tangles in brain parenchyma [10]
  - Emerging evidence also implicates additional pathophysiological pathways: neuroinflammation, axonal disintegration, brain metabolic dysfunction [3]
- AD accounts for 60-70% of the ~55 million dementia cases across the globe (WHO)
  - "The public health crisis of the 21st century" [7]



My Inspiration

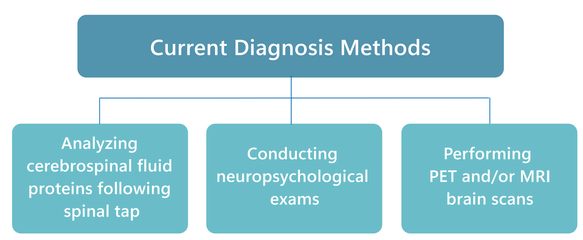


Photograph taken by Mahesh Sreenivas, Aug. 2021, in Trivandrum, Kerala, India.

My grandma—who used to be one of the most lively and talented people I knew—was unfortunately diagnosed with Alzheimer's Disease a couple of years ago, and that too, very late.

Visiting her in India this past summer was shocking to me in terms of how severely her condition had deteriorated within a span of just several months. Today, she no longer recognizes me or even my dad, her own son.

Current Situation & The Need



Diagnosis today is *inaccessible, invasive, expensive, and time-consuming*, involving a combination of the above three procedures.

*There is a **dire need** for globally accessible, affordable, and non-invasive methods to differentially and timely diagnose Alzheimer's.*

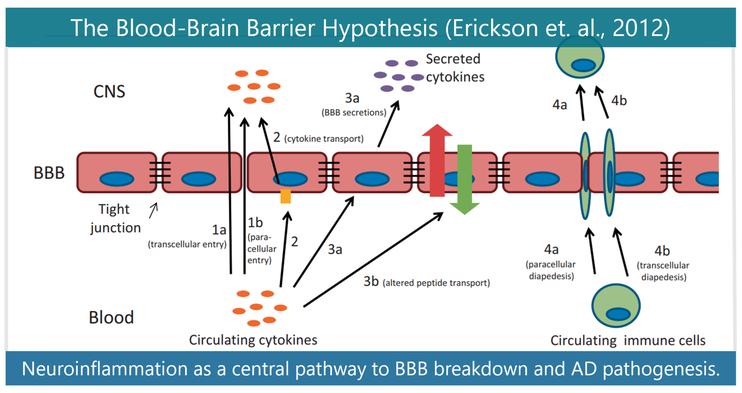
Identifying & validating a blood-based biomarker for AD would be groundbreaking since it would allow for routine & convenient monitoring of disease progression in patients with a simple blood test, facilitating earlier diagnoses.

Blood-Based Biomarkers

What do we already know?

The Blood-Brain Barrier (BBB) Hypothesis: the BBB becomes increasingly permeable with aging and neuroinflammation

- Effects: modification of transporter cells located near the BBB, altered interactions between cells of central nervous system (CNS) and bloodstream (e.g. immune cell trafficking; cytokine transport) [2]
- Significance: changes that occur up in the brain in AD patients may be detectable in the blood



Challenges & Limitations

- Heterogeneity and multi-layered complexity of Alzheimer's
- Relatively small study sample sizes + imbalnce in sex/social factors
- Clinical similarity to other non-AD dementias
- Frequent presence of co-morbidities (unaccounted for)

Why is there no validated blood-based biomarker for AD today?

The Problem: lack of reproducibility & consistency across studies

Current candidates: plasma Aβ1–42 and Aβ1–40, tau, neurofilament light, β-secretase 1 [3]

Can robust, reliable blood-based biomarkers unique to Alzheimer's Disease be identified by implementing machine learning techniques to analyze aggregated metabolomic and transcriptomic datasets?

Research Question & Goals

- Visualize the aggregated metabolomic and transcriptomic datasets to understand their features and limitations
- Develop various machine learning classification models to differentiate between Alzheimer's Disease & cognitively normal patients
- Analyze the best-performing machine learning model(s) to identify candidate robust blood-based biomarkers
- Perform gene ontology enrichment analysis to determine affected biological pathways based on the differentially expressed genes found

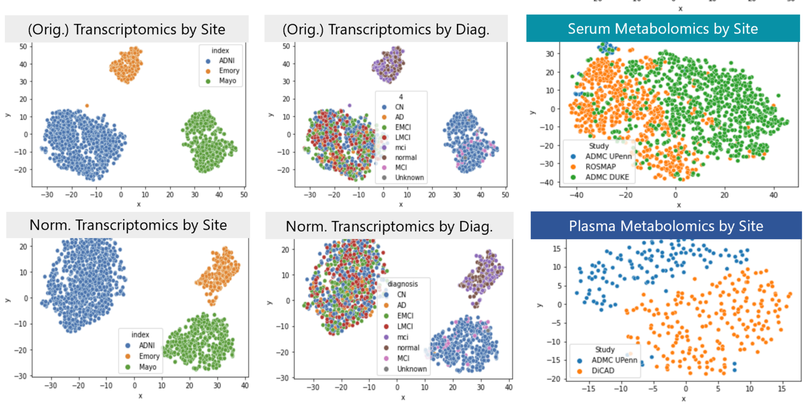
Methodology (Engineering Goals 1 & 2 ☑)

Data Aggregation

- Blood profiles of different study sites obtained from AD-related databases and by reaching out to study authors
- *Data pre-processing:* metadata compilation, concatenation & merging, data imputation
- Aggregated Dataset Dimensions
  1) Transcriptomics: 1358 samples, 17733 genes, 4 studies
  2) Serum Metabolomics: 1450 samples, 153 metabolites, 3 studies
  3) Plasma Metabolomics: 326 samples, 168 metabolites, 2 studies

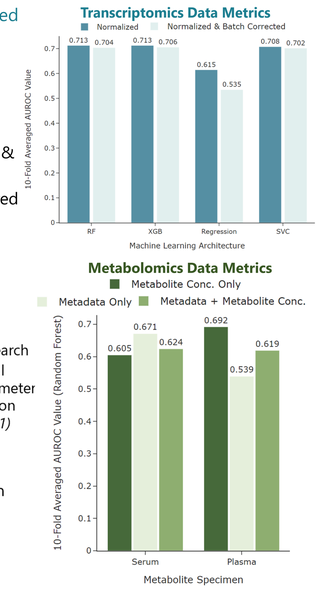
Data Visualization

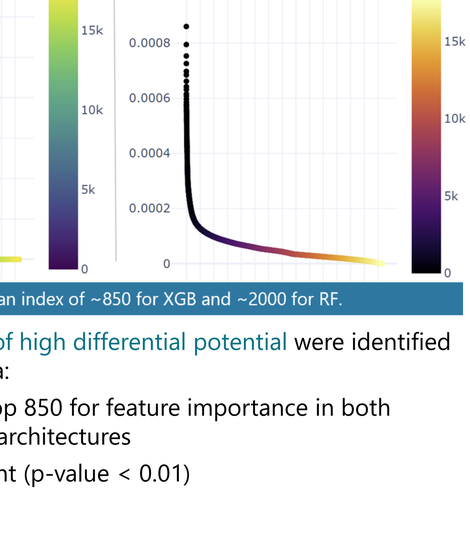
- Purpose: to understand the features and limitations of the datasets & evaluate whether transformations are necessary
- t-SNE algorithm used for dimensionality reduction & 2D visualization



Machine Learning

- Stratified 10-Fold Nested Cross-Validation Scheme to estimate an unbiased model generalization performance
- 4 Model Types: Random Forest (RF), Extreme Gradient Boosting (XGB), Support Vector Machine (SVM), Logistic Regression
- Conclusions
  - Transcriptomics showed the greatest potential as robust AD biomarkers & serum metabolites showed very little potential
  - The diagnostic potential of plasma metabolites should be further explored





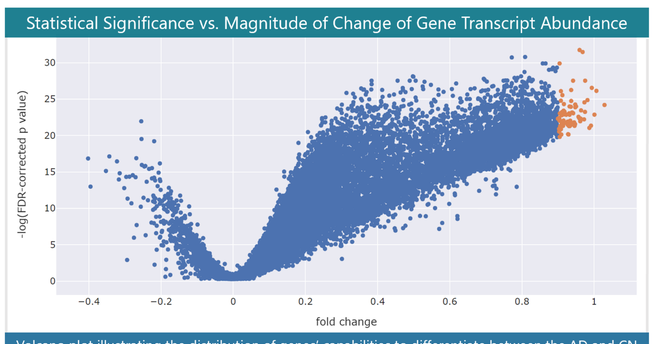
Data Analysis & Results (Engineering Goal 3 ☑)

Statistical Analyses

Differential potential of each of the genes in the transcriptomics dataset was assessed via:

- Mann-Whitney U Test with multiple test FDR corrections for calculation of p-value
- Fold change = (*mean blood transcript abundance of AD patients – mean of CN patients) / mean of CN patients*

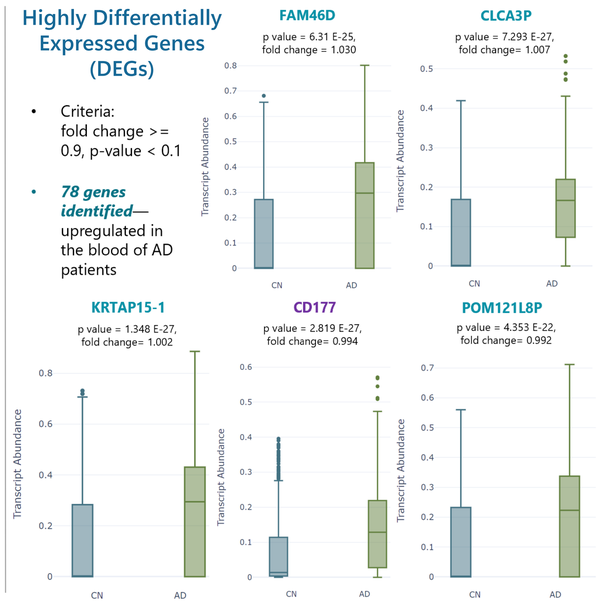
Statistical Significance vs. Magnitude of Change of Gene Transcript Abundance



Volcano plot illustrating the distribution of genes' capabilities to differentiate between the AD and CN study cohorts based on p-value and fold change. Orange-colored points indicate highly differentially expressed genes with a fold change >= 0.9 and a p-value < 0.01.

Highly Differentially Expressed Genes (DEGs)

- Criteria: fold change >= 0.9, p-value < 0.1
- ***78 genes identified***— upregulated in the blood of AD patients

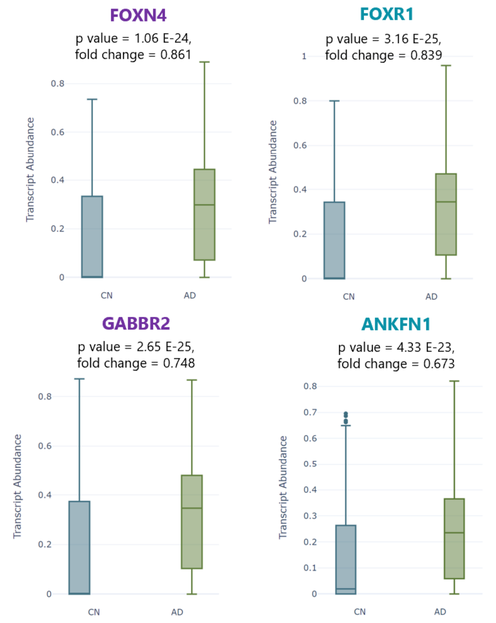


Machine Learning Feature Importance Analyses

Elbows located at an index of ~850 for XGB and ~2000 for RF.

An additional 9 genes of high differential potential were identified by the following criteria:

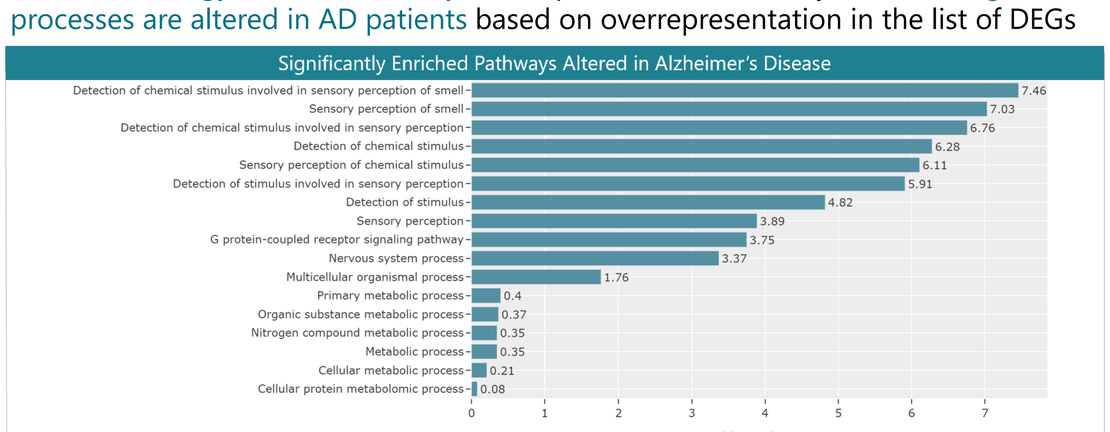
- Ranked within the top 850 for feature importance in both top-performing ML architectures
- Statistically significant (p-value < 0.01)
- Fold change > 0.5



Discussion & Conclusions (Engineering Goal 4 ☑)

Gene ontology enrichment analysis was performed to identify which biological processes are altered in AD patients based on overrepresentation in the list of DEGs

Significantly Enriched Pathways Altered in Alzheimer's Disease



Fold enrichment value represents the fold increase in the frequency genes of a given pathway were present in the inputted DEGs list relative to what would be expected in a random gene list. All plotted pathways are statistically significant with p-values < 0.01. In this context, the p-value represents the chance that the observed frequency of genes mapped to the specific term is due to random chance.

- Successfully identified 87 genes whose blood transcript abundances show promising differentiating capability between Alzheimer's and cognitively healthy patients *across multiple study sites*
  - ***Significance:*** the genes represent candidate robust blood-based biomarkers that may take us a step closer towards one day diagnosing Alzheimer's with a simple blood test
- These genes were enriched with various biological pathways that are characteristic of AD pathogenesis
  - ***Significance:*** corroborates the Blood-Brain Barrier Hypothesis and represents potential therapeutic targets to advance the quest for a cure
- Serum metabolites showed little to no potential as Alzheimer's biomarkers
- Limitations
  - Control cohort was limited to cognitively healthy patients only
  - Presence of co-morbidities were unaccounted for

Future Opportunities

- Validate that a subset of the DEGs identified also have potential in differentiating AD from other types of dementias
- Further investigate the novel DEGs to determine whether they are correlated with any comorbidities
- Extend study to other multiomics fields such as proteomics
- Conduct further research on plasma metabolites