Transcriptomic Biomarkers for Early Alzheimer's Detection through Machine Learning

Abbas Mirza
Joie Chen

Narayanan, M. *et al.* Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. *Mol. Syst. Biol.* **10**, 743. doi:10.15252/msb.20145304 (25 July, 2014)

Liang, W.S. *et al.* Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: a reference data set. *Physiol. Genomics* **33**, 240–256. doi:10.1152/physiolgenomics.00248.2007 (5 February, 2008)

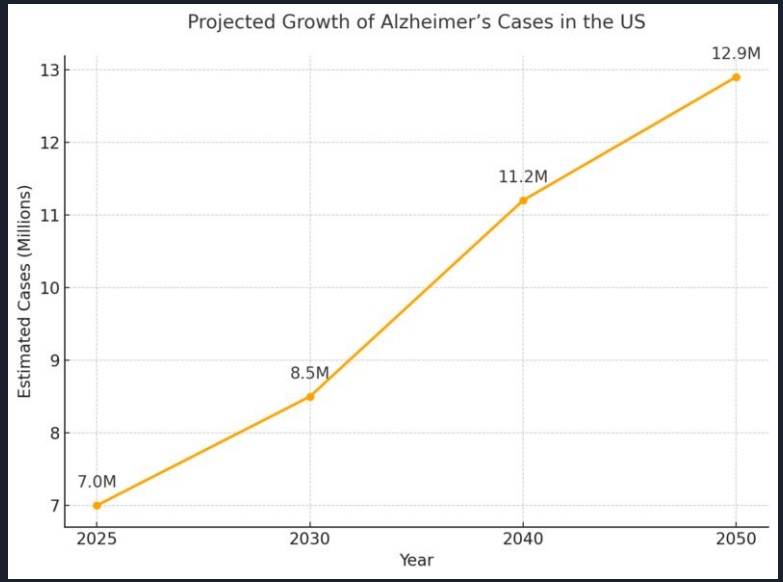
Cribbs, D.H. *et al.* Extensive innate immune gene activation accompanies brain aging, increasing vulnerability to cognitive decline and neurodegeneration: a microarray study. *J. Neuroinflammation* **9**, 179. doi:10.1186/1742-2094-9-179 (25 July, 2012)

Introduction

Problem Description

Alzheimer's Disease (AD) is the leading cause of dementia,

- Affecting 7.2 million Americans aged 65+ (2025); projected to reach 13.8 million by 2060
- 6th leading cause of death in U.S. adults
- Annual care costs: $384 billion, projected to exceed $1 trillion by 2050
- Diagnosis is often delayed by years; early symptoms missed or misattributed
- Urgent need for early, reliable, and scalable diagnostic tools



Early detection can:

- Allow timely intervention
- Improve quality of life
- Enable inclusion in clinical trials for emerging therapies

Biology/Physiology

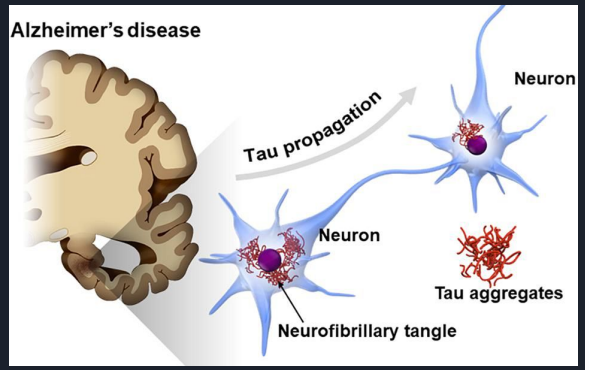


Image adapted from Dong et al., *Frontiers in Neuroscience* (2019), https://doi.org/10.3389/fnins.2019.01274.

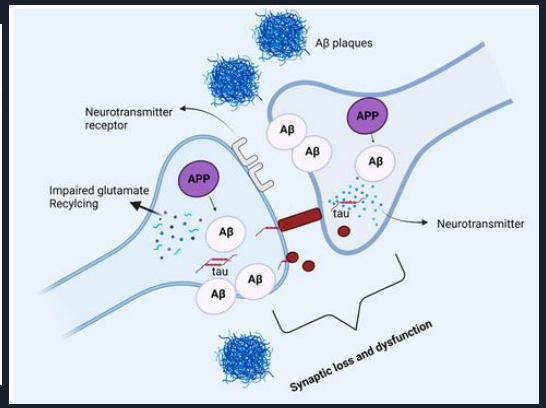


Image adapted from Raimondi et al., *International Journal of Molecular Sciences* (2022), https://doi.org/10.3390/ijms232112924.

- Pathological Changes:
  - β-amyloid plaque buildup (extracellular)
  - Tau protein tangles (intracellular)
- Leads to synaptic dysfunction, neuroinflammation, and neuronal loss
- Affects key regions: prefrontal cortex, hippocampus, entorhinal cortex
- Molecular changes begin 10–20 years before cognitive decline
- Transcriptomic changes provide early biomarkers for detection

Goals

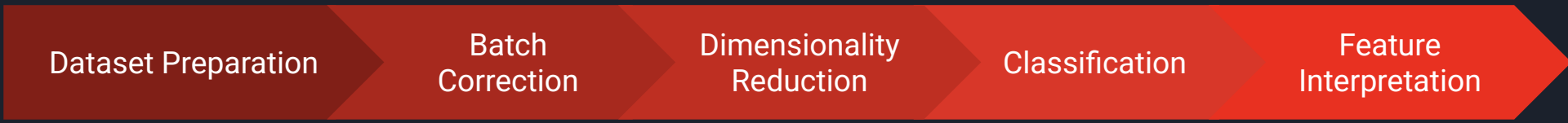
- Use ML to identify gene expression biomarkers for early AD detection
- Assess model performance and interpret key features (genes) contributing to prediction.
- If Successful:
  - Could lead to non-invasive diagnostics using gene expression profiling.
  - Contributes to precision medicine approaches in neurology.

Datasets

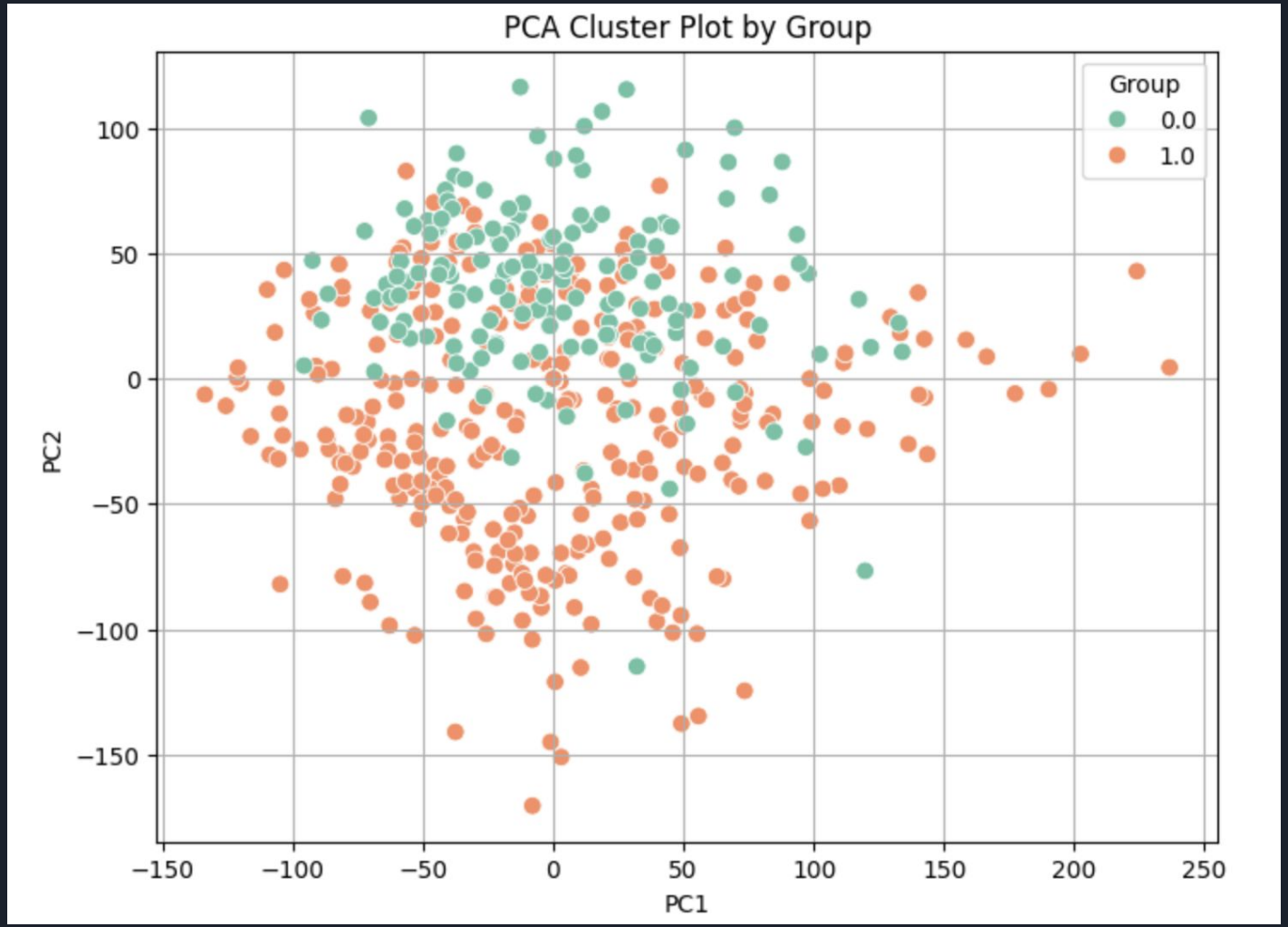
- GSE33000
  - 624 DLPFC (BA9) samples: Alzheimer's, Huntington's, and controls.
  - Collected post-mortem; profiled using Agilent 44K array (GPL4372).
  - Source: Harvard Brain Tissue Resource Center (HBTRC).
  - No normalization
- GSE5281
  - Samples from 6 brain regions across AD progression.
  - Tissue processed with laser capture microscopy to reduce heterogeneity.
  - Affymetrix U133 Plus 2.0 platform (~55,000 transcripts).
- GSE48350
  - 4 regions: hippocampus, entorhinal cortex, superior frontal cortex, post-central gyrus.
  - Includes normal controls (ages 20–99) and AD patients.
  - Investigates both aging-related and AD-specific gene expression changes.
- Note:
  - Preprocessing included standard microarray normalization (RMA, quantile) and quality control per each dataset's original study.

Methods

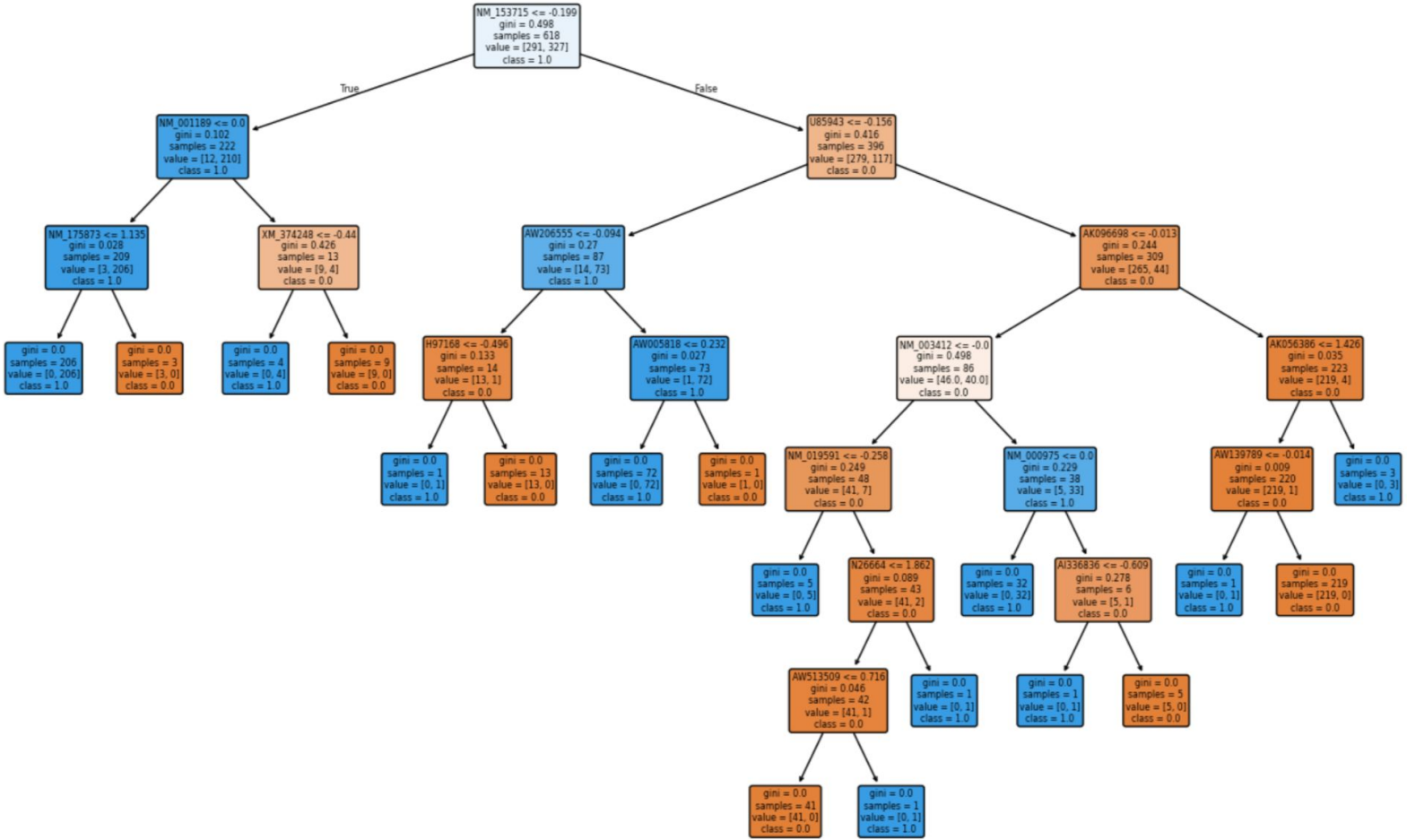
- Programing Language: Python
- Normalization was checked per dataset by checking mean and median.
- Z-score normalization applied within each dataset thereafter
- Batch correction was performed using the ComBat method to minimize platform-related effects
- Used Principal Component Analysis (PCA) to visualize sample clustering and detect batch effects
- Logistic Regression and Decision Tree classifiers were trained to distinguish AD vs. control samples.
- Examined coefficients from the logistic regression model and use feature importance scores from the decision tree.

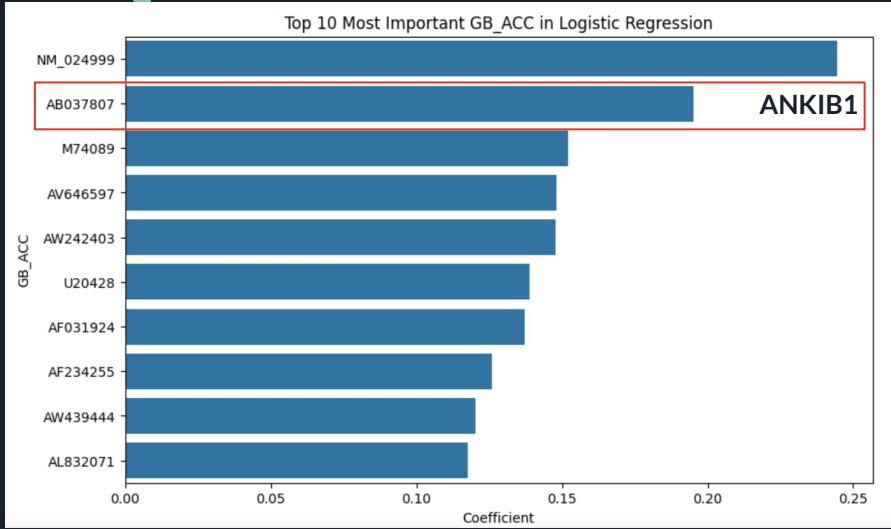


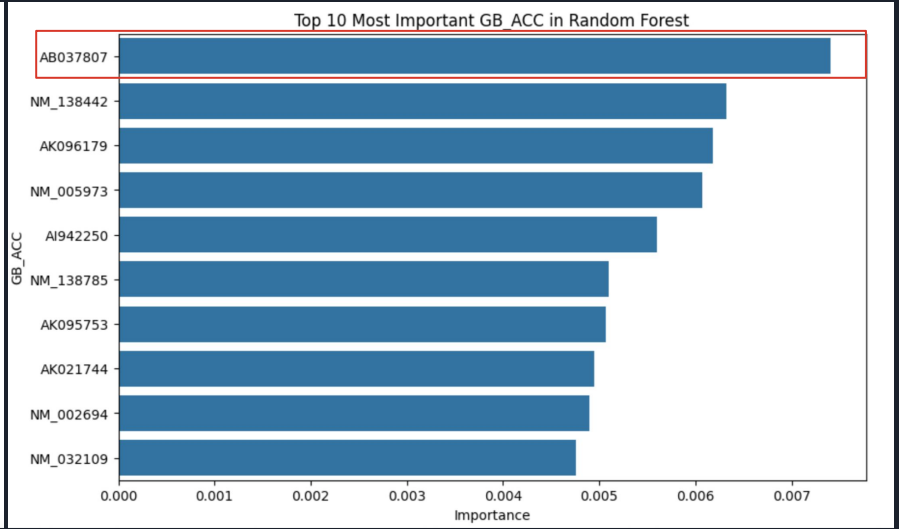
Results


PCA Cluster Plot by Group
Group
0.0
1.0
PC2
PC1
100
50
0
−50
−100
−150
−150
−100
−50
0
50
100
150
200
250

Decision Tree Visualization







Logistic Regression	Linear classifier on expression data	**Accuracy: 94%**
Random Forest	Nonlinear ensemble model	**Accuracy: 88%**

Discussion

Biological Relevance

- **ABCA7**: Implicated in lipid transport & amyloid-β processing, linked to AD risk. (Hollingworth et al., *Nat Genet*, 2011)
- **PLD3**: Affects lysosomal function and amyloid metabolism. (Cruchaga et al., *Nat Neurosci*, 2014)
- **FAM3C**: Downregulated in AD, may support synaptic function. (Gupta et al., *J Alzheimers Dis*, 2016)
- **FAM222A**: Found in amyloid plaques, potential role in AD pathology. (Wang et al., *Nature*, 2021)

Limitations

- Residual batch effects may remain.
- Some probe-to-gene mappings may be ambiguous.
- No longitudinal data; only cross-sectional.

Future Directions

- Validate top genes in external datasets.
- Explore gene function in neuronal models.
- Use longitudinal data to track disease progression.
- Integrate with proteomics/metabolomics for deeper insight.

References

- Alzheimer's Association. (2025). *2025 Alzheimer's Disease Facts and Figures*. Retrieved from https://www.alz.org/alzheimers-dementia/facts-figures
- Centers for Disease Control and Prevention. (2024). *About Alzheimer's Disease*. Retrieved from https://www.cdc.gov/alzheimers-dementia/about/alzheimers.html
- Mayo Clinic. (2024). *Alzheimer's disease - Symptoms and causes*. Retrieved from https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447
- Jack, C. R., et al. (2010). *Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade*. *The Lancet Neurology*, 9(1), 119–128.
- Wang, M., et al. (2019). *Peripheral transcriptomic biomarkers for early detection of sporadic Alzheimer's disease*. *Nature Communications*, 10, 1–13. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6436957
- Selkoe, D. J., & Hardy, J. (2016). *The amyloid hypothesis of Alzheimer's disease at 25 years*. *EMBO Molecular Medicine*, 8(6), 595–608.
- Warner Senate. (2025). *Warner, Capito Introduce Bill to Improve Early Assessment & Diagnosis of Alzheimer's*. Retrieved from https://www.warner.senate.gov

Thank You!