Final Presentation:
Alzheimer’s disease neuropathology prediction

Seo-Yoon Moon

CSE428
2023.06.01

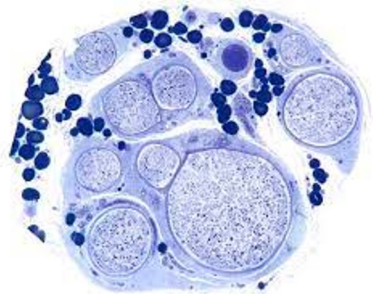
Motivation and background

Alzheimer's disease (AD) is a well-known disease in the neuroimaging field, but associations between genes and AD is not fully uncovered. Also, if it possible to predict neuropathology in AD patients, clinicians can provide the treatment targeting the specific neuropathology.

*Neuropathology: the study of diseases of the brain, spinal cord, and nerves through analyzing tissues

People often use six pathologies:

(1) Aβ IHC: amyloid-β protein density via immunohistochemistry
(2) plaques: neuritic amyloid plaque counts from stained slides
(**3) CERAD** score: a semi-quantitative measure of neuritic plaque severity
(4) τ IHC: abnormally phosphorylated τ protein density via immunohistochemistry
(5) tangles: neurofibrillary tangle counts from silver-stained slides
(6) Braak stage: a semi-quantitative measure of neurofibrillary tangle pathology



Problem setting

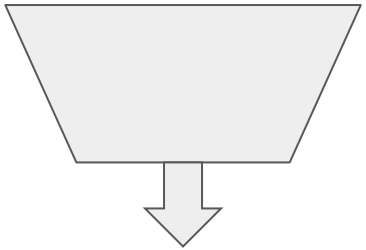
Problem

Predicting neuropathology (CERAD) of Alzheimer's disease

Formal definition

- **Input:** Gene expression data & description of each gene from ChatGPT
- **Output:** Level of neuropathology (CERAD) - numerical value

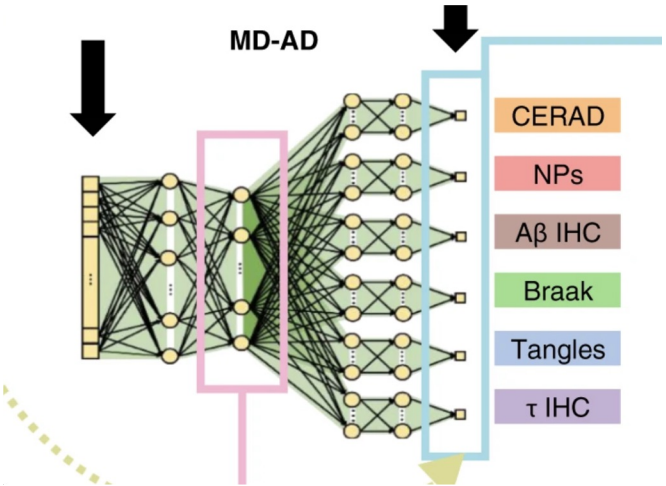
Gene expression data
+
chatGPT description



Level of neuropathology (CERAD)
(e.g. 1.0, 3.0, 5.0 …)

Baseline

MD-AD model



Model summary

- Used only gene expression data for inputs
- Gene expression data is PCA transformed
- Investigated simple MLP model, ML models are not tested

Implementation

- Code are provided: https://github.com/suinleelab/MD-AD

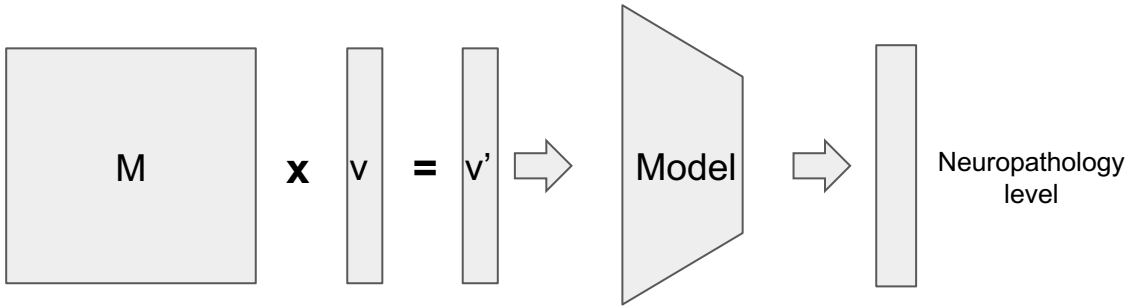
Result

- In the paper: 0.44 (R-squared)
- My reproduction: -0.31 (R-squared)

Beebe-Wang, N., Celik, S., Weinberger, E. *et al.* Unified AI framework to uncover deep interrelationships between gene expression and Alzheimer's disease neuropathologies. *Nat Commun* 12, 5369 (2021). https://doi.org/10.1038/s41467-021-25680-7

Our method

1. Select 200 genes that are associated with neurofibrillary tangles (NFT) in Alzheimer's disease
   a. ChatGPT
   b. Bard
   c. MD-AD paper
2. Generate description about how they are involved in forming NFT by asking ChatGPT
3. Calculate embedding and similarity using sentence transformer
4. Put into the random forest model and get prediction of outcome.

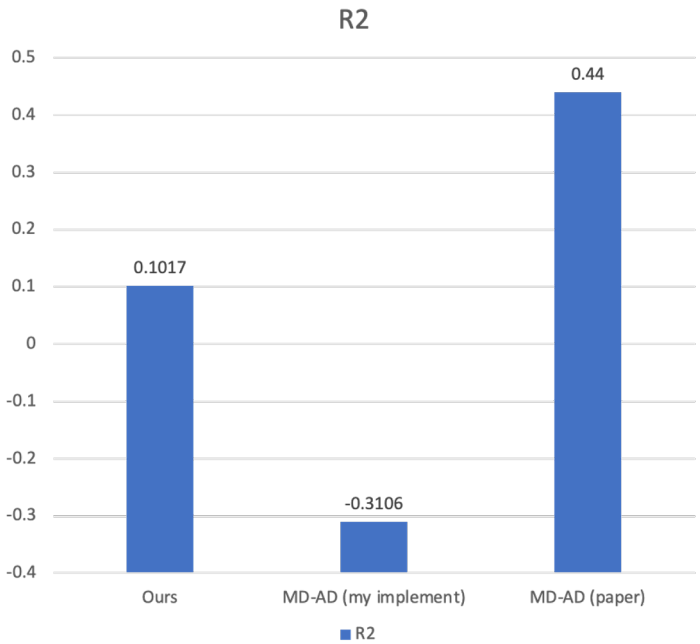


M: similarity matrix of description, v: gene expression

Why ChatGPT are used?

- When selecting the genes, it is better to choose genes that have already been shown to be involved.
- Since ChatGPT can summarize all the associations that it knows, so I used for generating description.

Results



Question to ChatGPT

What is the gene {g}'s role and how is it related to neurofibrillary tangle, one of the Alzheimer's disease pathology?

Possible reasons that our method is not better than MD-AD

- I only used small amount of genes
- ChatGPT often didn't generate description based on real paper.
- MD-AD has two steps: the first step is for training shared features across all neuropathologies in Alzheimer's disease and the second step is for training neuropathology-specific features.
- My model only focuses on one neuropathology. Since the pathologies affect each other, training shared features could be a good choice.
- I also trained MLP but the result became worse.
- I also tried adding similarity between pathology description and gene description - but the result didn't change.

Conclusion and future work

- ChatGPT often generate wrong descriptions: given the same prompt, the positive/negative effect sometimes changed.
- If given gene has any association with Alzheimer's disease, ChatGPT sometimes randomly creates the associativity.
- ChatGPT was not very useful on this dataset. But I think if we use small datasets which are lack of information.

If I have more time, I would like to

- Explore how many % correct descriptions chat gpt generates.
- With a small number of genes that are approved to be significantly associated with Alzheimer's disease, I want to train models again.