Final Report: Transcriptomic Biomarkers for Predicting the Progression from Mild Cognitive Impairment to Alzheimer's Dementia.

Members: Kitiyaporn Takham, Mahitha Chaturvedula

Code: https://github.com/Kitiyaparnnn/MCI-AD.git

1. Introduction

Alzheimer's disease (AD) is the *leading cause of dementia*, with roughly 80% of cases around the world arising from AD [1]. However, effective treatment for AD remains an unmet need. Mild cognitive impairment (MCI) is the initial phase of memory decline for patients and is often underdiagnosed due to its confusion with AD [2]. In the last 20 years, mild cognitive impairment (MCI) has been said to designate an early, yet abnormal, state of cognitive impairment [3] and has been classified as a high-risk factor for the development of AD [4]. Therefore, it is critical to identify patients with MCI who are at high risk of progressing to AD early on. Clearly defining expression markers for MCI and AD might lead to earlier detection and medical care for dementia patients.

Many studies have suggested that biomarkers obtained by comparing AD to controls can be used to predict conversion from MCI to AD. However, recent studies have indicated that this change has a non-linear trajectory where cognitively unimpaired (CU) individuals progress to MCI, and then further progress to AD [4,5]. A recent imaging study found that molecular biomarkers predicting CU-to-MCI conversion are not as helpful as they are for MCI-to-AD conversion [6]. This is why we aim to identify specific biomarkers to help characterize this MCI-to-AD conversion.

This project has two aims: first, to identify a set of gene expression biomarkers for high-risk and low-risk MCI to AD dementia prognosis, and second, to explore the potential of leveraging an AI classifier by comparing the performance of supervised models between machine learning and deep learning approaches.

2. Data

Dataset	Modality & Data Type	Role in Project	Access Information
GEO: GSE282742 [7]	Variables: Age, Sex, and Gene expression of white blood cells (n=61,860) - progressive MCI(P-MCI), n=28 - stable MCI (S-MCI), n=39 - AD, n=49	Training Cohort	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE282742
GEO: GSE249477 [8]	Variables: Age, Sex, and Gene expression of white blood cells (n=21,492) - MCI due to AD, n=20 - AD, n=21	External Validation Cohort	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE249477

3. Methods

Within the data preprocessing process, to ensure compatibility between the training cohort and the external cohort, genes were intersected between the cohorts, resulting in 21,462 shared genes. All gene expression values underwent normalization by log2(TMP + 1) across the gene expression and filtering for high-information genes by using the mutual information method to collect 2,000 high-informative genes. Link to processed train data and external data.

Based on the MCI training data, we considered 2 task classifications:

(1) MCI vs. AD classification: we combined S-MCI and P-MCI together and trained them as an MCI to identify biomarkers to classify between MCI and AD.
(2) S-MCI vs. P-MCI vs. AD classification: to identify biomarkers to classify between S-MCI, P-MCI, and AD.

We implemented two supervised learning methods, SVM – and Logistic regression, and 1D-CNN, to compare their classification performance and identify the gene biomarkers for each task.

3.1. Support Vector Machine (SVM) and Logistic Regression (LR)

First, we implemented GridSearch with 5-fold cross-validation on each of the SVM and LR models (LinearSVC, LogisticRegression, respectively), setting a maximum of 5000 iterations and using balanced weights to address class imbalance, in order to find the optimal C index based on accuracy. We then used the best C index to identify the optimal gene set through the leave-one-out (LOOCV) method with L2 regularization among 200 genes, which are common gene features for SVM training. We looped this process for each task.

3.2. Convolutional Neural Network (CNN)

The 1D-CNN model contains 3 convolutional layers (32, 64, 128): BatchNorm, ReLu activation function, and AdaptiveMaxPooling, then 0.3 Dropout and one fully connected layer (128) using the PyTorch library. First, we found the optimal gene set among the initial 2000 genes by using the elbow point of the integrated gradient (IG) score that was trained on stratified 5-fold cross-validation with a weighted balance cross-entropy loss function (to solve imbalance classes), Adam optimizer, 50 epochs, and 0.001 learning rate. Then we trained the final model with 5-fold cross-validation and the optimal gene set for each task.

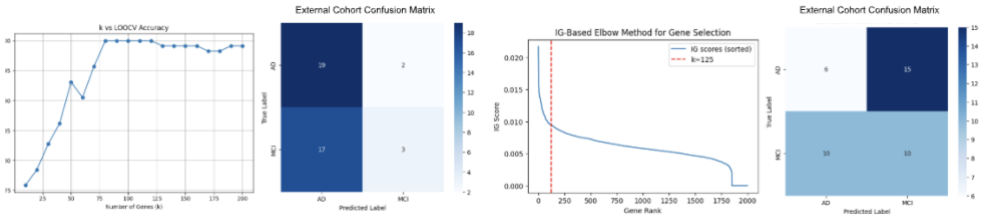
3.3. Evaluation

We evaluated the model's performance using accuracy, specificity, and sensitivity matrices for each task shown below:

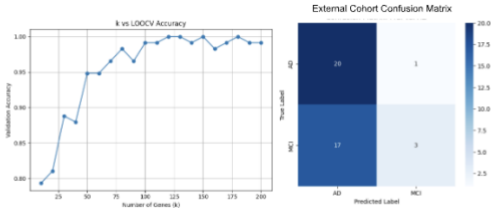
(1) MCI vs. AD classification: the final model with the optimal gen set was used to directly evaluate the external cohort to classify between MCI and AD.
(2) S-MCI vs. P-MCI vs. AD classification: the final model with the optimal gene set was used to classify the external cohort; however, the prediction was converted to both S-MCI and P-MCI to MCI before calculating the matrix.

In terms of gene biomarkers, we compared the optimal gene set from each model per task using GSEA to identify enriched pathways and gene-level interpretation.

4. Results



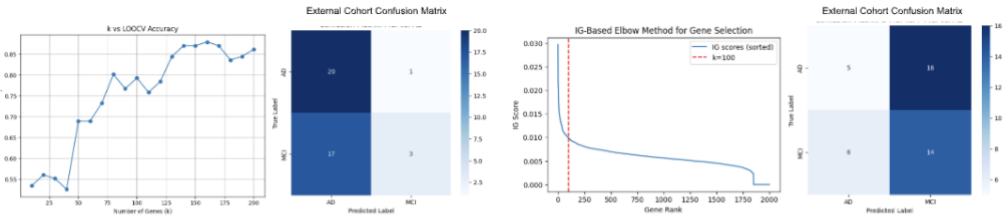
SVM Model 1D-CNN Model



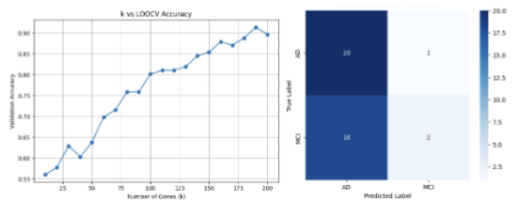
LR Model

For MCI vs. AD classification, the SVM with the 80 optimal genes achieved 99.14% average accuracy, 97.96% average specificity, and 100% average sensitivity to MCI on the training cohort. Similar to the LR model with 120 optimal genes, it performed 100% average accuracy, 100% average sensitivity, and specificity for MCI. However, they didn't perform well on the external cohort with 53.66% (SVM) and 56.10% (LR) accuracy, and 15.00% sensitivity and specificity to MCI on both models.

As for the 1D-CNN model with 125 genes (considered as the elbow point on the minimum genes to reduce the computational cost and time-consuming issue), the model reached only 59.30% average accuracy, 83.30% average sensitivity to MCI, and 60.00% specificity to MCI on the training cohort. Similarly, the CNN model performed worse on the external cohort, with 39.02% average accuracy, and was really biased to the MCI class.



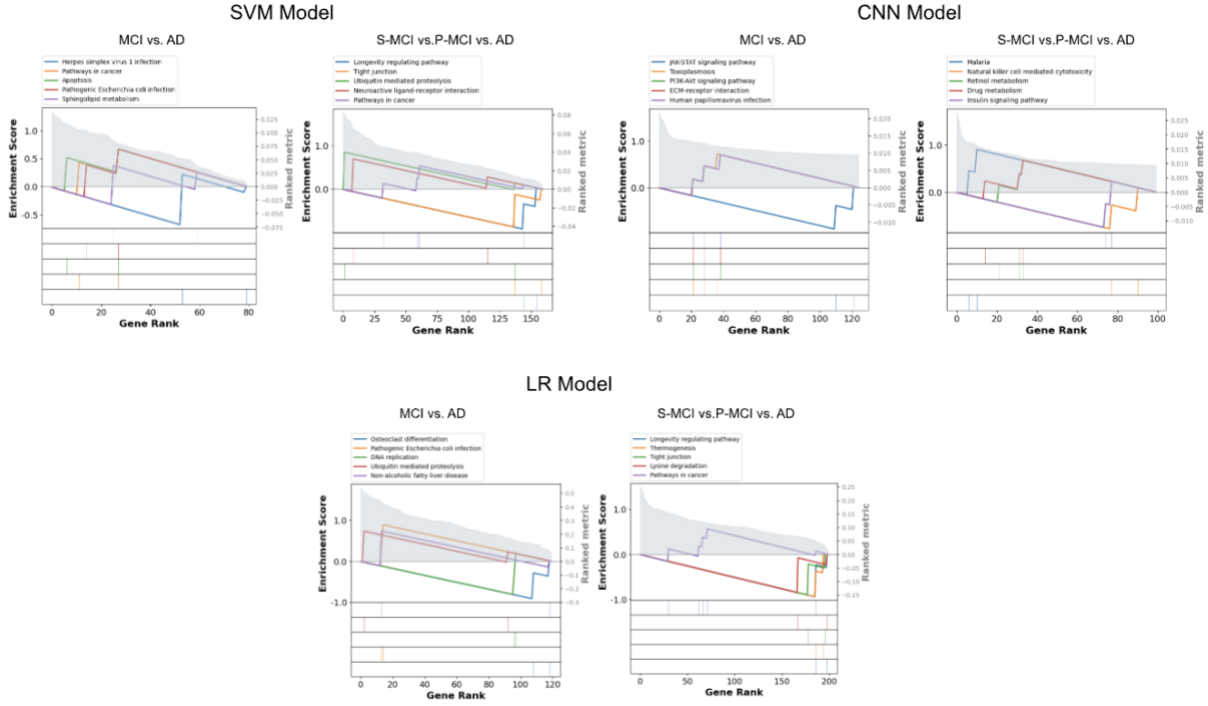
SVM Model 1D-CNN Model



LR Model

For S-MCI vs. P-MCI vs. AD classification, the SVM with 160 optimal genes achieved 82.80% accuracy, average specificity (S-MCI: 84.21%, P-MCI: 92.05%, AD: 98.51%), and average sensitivity (S-MCI: 84.62%, P-MCI: 67.86%, AD: 89.80%) on the training cohort. On the external evaluation, the SVM's performance dramatically dropped to 56.10% accuracy, and bias was diagnosed to AD with 95.23% MCI specificity and 15.00% MCI sensitivity. In contrast, the LR with 190 optimal genes outperformed on the training cohort with 90.5% accuracy, average specificity (S-MCI: 92.21%, P-MCI: 96.59%, AD: 100.00%), and average sensitivity (S-MCI: 92.31%, P-MCI: 82.14%, AD: 97.96%), but the performance on the external cohort was worse with 53.66% accuracy, almost predicted as AD (95.24% MCI specificity and 0.10% MCI sensitivity).

With 100 optimal genes on a 1D-CNN model, it performed with 42.14% average accuracy, average sensitivity (S-MCI: 39.96%, P-MCI: 19.52%, AD: 57.72%), and average specificity (S-MCI: 54.77%, P-MCI: 92.33%, AD: 62.06%). The model's performance was not good on the external cohort as well; it achieved 36.59% accuracy.



Across the three models, GSEA reveals both overlapping and model-specific pathway patterns. SVM and LR show stronger consistency, with several immune-related and proteolysis pathways recurring in both MCI vs. AD and S-MCI vs. P-MCI vs. AD comparisons. CNN highlights more metabolism- and signaling-related pathways, suggesting that its feature ranking emphasizes different biological mechanisms than the linear models. Overall, while all models capture immune and infection-related signatures, the CNN diverges by prioritizing metabolic and signaling pathways, reflecting differences in how each model ranks gene importance.

We found potential MCI/AD-related biomarker genes indicated similarly in both SVM and LR gene sets, including CASP7 (caspase-7): implicated in neuronal apoptosis and AD risk [9],

COL4A1 (basement-membrane/vascular gene linked to blood–brain barrier integrity): implicated in early cognitive decline and conversion to AD [10], GLB1 (lysosomal/glycosphingolipid metabolism and senescence markers): altered in AD and associated with progression [11], and PPARG (PPAR-γ modulates inflammation and metabolism): influences pathways that could slow progression [12]. However, with the different domains in CNN, it also includes potential biomarker genes such as PON1 (antioxidant/paraoxonase enzyme): lower PON1 activity has been associated with MCI and with conversion risk [13] and CXCL8 (pro-inflammatory chemokine): associated with faster cognitive decline [14].

5. Conclusion

The LR model outperformed the SVM and 1D-CNN models on training and external cohorts in overall accuracy, sensitivity, and specificity for both MCI vs. AD and S-MCI vs. P-MCI vs. AD, even though none of them reached higher than 60% accuracy on the external cohort. Due to the lack of stable and progressive MCI gene expression observed in the training cohort, there is a limitation in evaluating the model's performance for classifying MCI patients and AD patients in the external cohort, as we must consider whether stable MCI or progressive MCI should be identified as MCI patients or classified as AD patients.

Nevertheless, the important gene sets from the three models include the significant MCI/AD conversion genes, such as CASP7, COL4A1, GLB1, PPARG, PON1, and CXCL8, which means that these approaches can be used to identify gene biomarkers. Also, the rest of the gene set is suggestive or indirect evidence of the MCI/AD conversion.

6. References

1. Gauthier S, Reisberg B, Zaudig M, Petersen RC, Ritchie K, Broich K, et al. Mild cognitive impairment. Lancet. 2006;367(9518):1262–1270. doi: 10.1016/S0140-6736(06)68542-5.
2. Mian M, Tahiri J, Eldin R, Altabaa M, Sehar U, Reddy PH. Overlooked cases of mild cognitive impairment: Implications to early Alzheimer's disease. Ageing Res Rev. 2024 Jul;98:102335. doi: 10.1016/j.arr.2024.102335. Epub 2024 May 12. PMID: 38744405; PMCID: PMC11180381.
3. Petersen RC. Mild cognitive impairment as a diagnostic entity. J Intern Med. 2004 Sep;256(3):183-94. doi: 10.1111/j.1365-2796.2004.01388.x. PMID: 15324362.
4. Gaubert S, Raimondo F, Houot M, Corsi MC, Naccache L, Diego Sitt J, et al. EEG evidence of compensatory mechanisms in preclinical Alzheimer's disease. Brain. 2019;142(7):2096–2112. doi: 10.1093/brain/awz150.
5. Di Costanzo A, Paris D, Melck D, Angiolillo A, Corso G, Maniscalco M, et al. Blood biomarkers indicate that the preclinical stages of Alzheimer’s disease present overlapping molecular features. Sci Rep. 2020;10:15612.
6. Li X, Wang H, Long J, Pan G, He T, Anichtchik O, et al. Systematic analysis and biomarker study for Alzheimer’s disease. Sci Rep. 2018;8:17394.
7. Huang YL, Tsai TH, Shen ZQ, Chan YH et al. Transcriptomic predictors of rapid progression from mild cognitive impairment to Alzheimer's disease. Alzheimers Res Ther 2025 Jan 3;17(1):3. PMID: 39754267
8. Iga JI, Yoshino Y, Ozaki T, Tachibana A, Kumon H, Funahashi Y, Mori H, Ueno M, Ozaki Y, Yamazaki K, Ochi S, Yamashita M, Ueno SI. Blood RNA transcripts show changes in inflammation and lipid metabolism in Alzheimer's disease and mitochondrial function in mild

cognitive impairment. J Alzheimers Dis Rep. 2024 Dec 23;8(1):1690-1703. doi: 10.1177/25424823241307878. PMID: 40034360; PMCID: PMC11863738.
9. Ayers KL, Mirshahi UL, Wardeh AH, et al. A loss of function variant in CASP7 protects against Alzheimer's disease in homozygous APOE ε4 allele carriers. BMC Genomics. 2016;17 Suppl 2(Suppl 2):445. Published 2016 Jun 23. doi:10.1186/s12864-016-2725-z
10. Plaisier E, Ronco P. COL4A1-Related Disorders. 2009 Jun 25 [Updated 2016 Jul 7]. In: Adam MP, Bick S, Mirzaa GM, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2025. Available from: https://www.ncbi.nlm.nih.gov/sites/books/NBK7046/
11. Fancy NN, Smith AM, Caramello A, et al. Characterisation of premature cell senescence in Alzheimer's disease using single nuclear transcriptomics. Acta Neuropathol. 2024;147(1):78. Published 2024 May 2. doi:10.1007/s00401-024-02727-9
12. Sato T, Hanyu H, Hirao K, Kanetaka H, Sakurai H, Iwamoto T. Efficacy of PPAR-γ agonist pioglitazone in mild Alzheimer disease. Neurobiol Aging. 2011;32(9):1626-1633. doi:10.1016/j.neurobiolaging.2009.10.009
13. Cervellati C, Trentini A, Romani A, et al. Serum paraoxonase and arylesterase activities of paraoxonase-1 (PON-1), mild cognitive impairment, and 2-year conversion to dementia: A pilot study. J Neurochem. 2015;135(2):395-401. doi:10.1111/jnc.13240
14. Zhou, F., Sun, Y., Xie, X. et al. Blood and CSF chemokines in Alzheimer’s disease and mild cognitive impairment: a systematic review and meta-analysis. Alz Res Therapy 15, 107 (2023). https://doi.org/10.1186/s13195-023-01254-1