Researching the Cell – Amyloid Plaque Relationship in Alzheimer's Disease

Author: Dimitar Smenovski [1] Responsible Professor: Marcel Reinders [1] Supervisors: Roy Lardenoije [1], Timo Verlaan [1], Gerard Bouland [1]

[1]EEMCS, Delft University of Technology, The Netherlands

1. Background

Alzheimer's disease (AD): A neuro-degenerative disease that is the leading cause of dementia in people above 65 years old [1].
Amyloid plaques: Toxic clumps of amyloid-β protein in the brain.
Gene expression: Transcription of DNA into proteins through mRNA molecules. Counts of each mRNA molecule within a cell are used to determine expression levels.
Single-cell RNA sequencing (scRNAseq): A technique that measures cells' gene expression by isolating them from the tissue.
Spatially resolved transcriptomics: A measuring technique, similar to scRNAseq, but which preserves cells' spatial features.

2. Introduction

ScRNAseq data has been used to train models that classify cells as AD [2]. However this data does not contain information on cells' location inside the tissue. Taking advantage of such information would allow to study the spatial properties of AD cells.

Amyloid plaques alter nearby cells. In particular, expression of plaque-induced genes is present in astrocytes, microglia and oligodendrocytes [3]. This leads to the following hypothesis:

> *Cells affected by AD should be located close to plaques, while heatlhy individuals (CT) should be further away.*

Spatial transcriptomic data measures fewer genes compared to scRNAseq data, due to lower precision. Adapting it to the single cell data is possible with integration methods [4] [5]. Then it can be used with AD classification models to study the hypothesis.

3. Methodology

Datasets: ROSMAP and SEA-AD (scRNAseq). Xenium (spatial).

Pre-process data:
Reduce scRNAseq data to 1000 most highly variable genes
Subset imputed spatial data to the 1000 training genes
Filter, normalize, logarithmize and scale all data
↓
Train Models on scRNAseq Data:
Linear Classification (LC), Multilayer Perceptron (MLP) and
MLP with a dropout layer (MLP+Dropout)
↓
Impute spatial data missing gene expressions:
Use Tangram for 5 epochs with both scRNAseq datasets
Validate using leave-one-gene-out cross-validation
↓
Classify imputed spatial data:
Use ROSMAP models on SEA-AD imputed data & vice versa
↓
Analyse:
Distance to plaque distribution by labels
(violin plots, Mann-Whitney U Test)
Correlation between AD probability and distance to plaque
(maps of cells' location within the tissue, Spearman's rho)

4a. Model and Validation Results

	LC	MLP	MLP+Dropout
ROSMAP	0.6984	0.9909	0.8840
SEA-AD	0.9225	1.0000	0.9997

Table 1. Final training accuracy per model and dataset

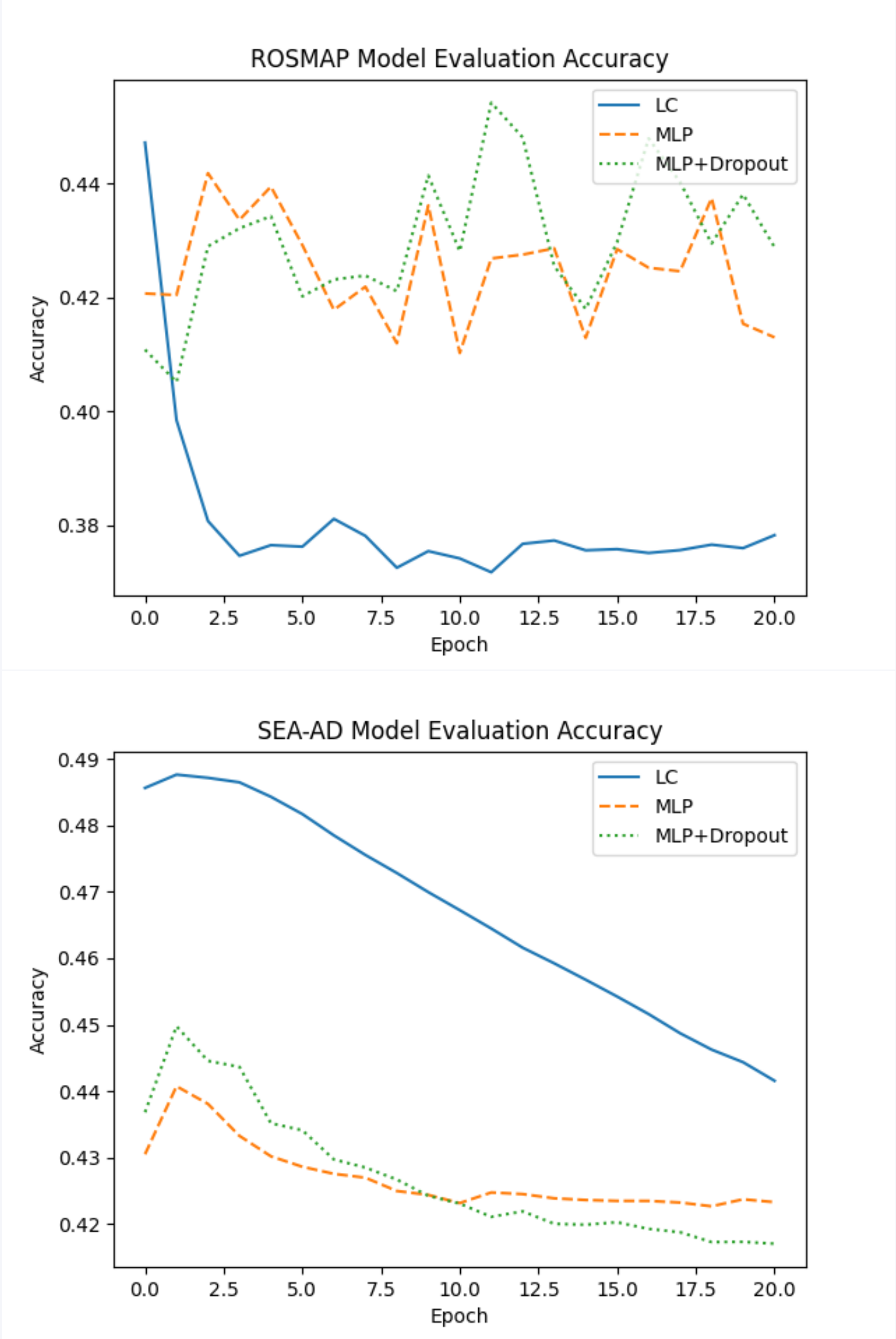

Figure 1. Evaluation accuracy per model and dataset

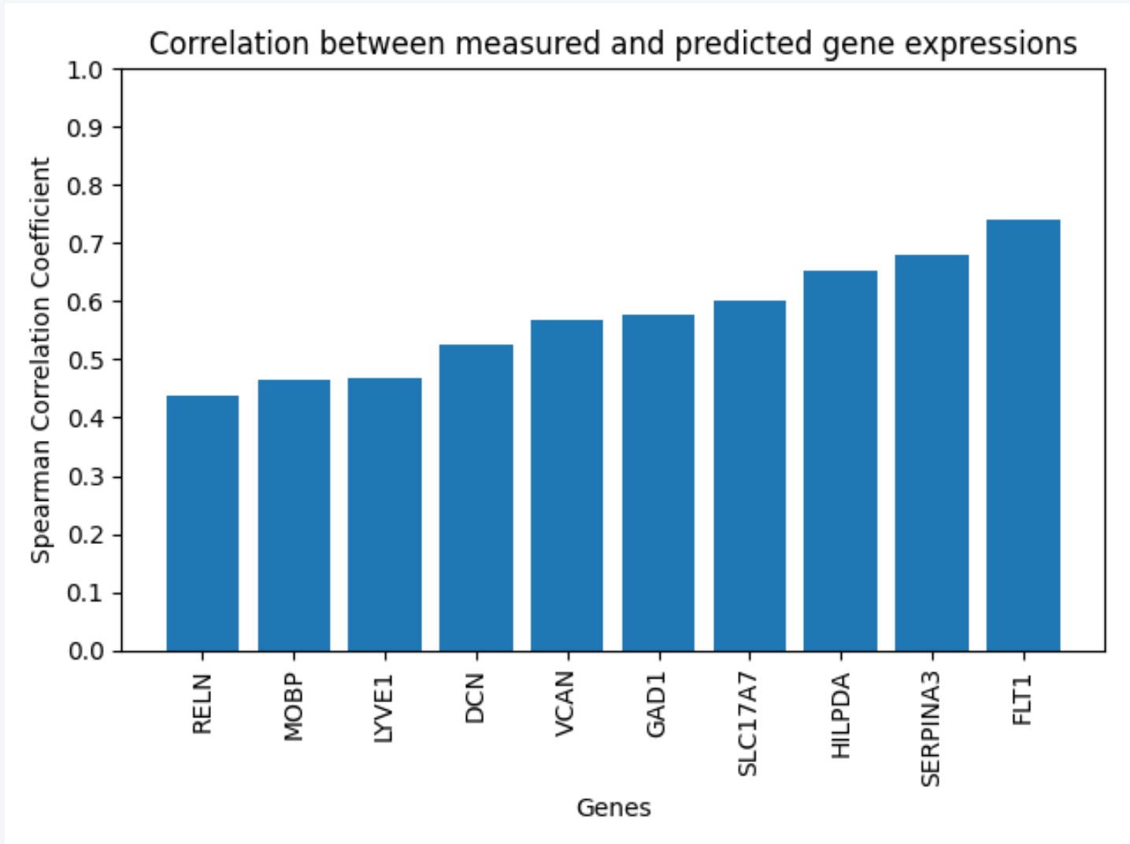

Figure 2. Correlation of imputed and true gene expression

4b. ROSMAP Results

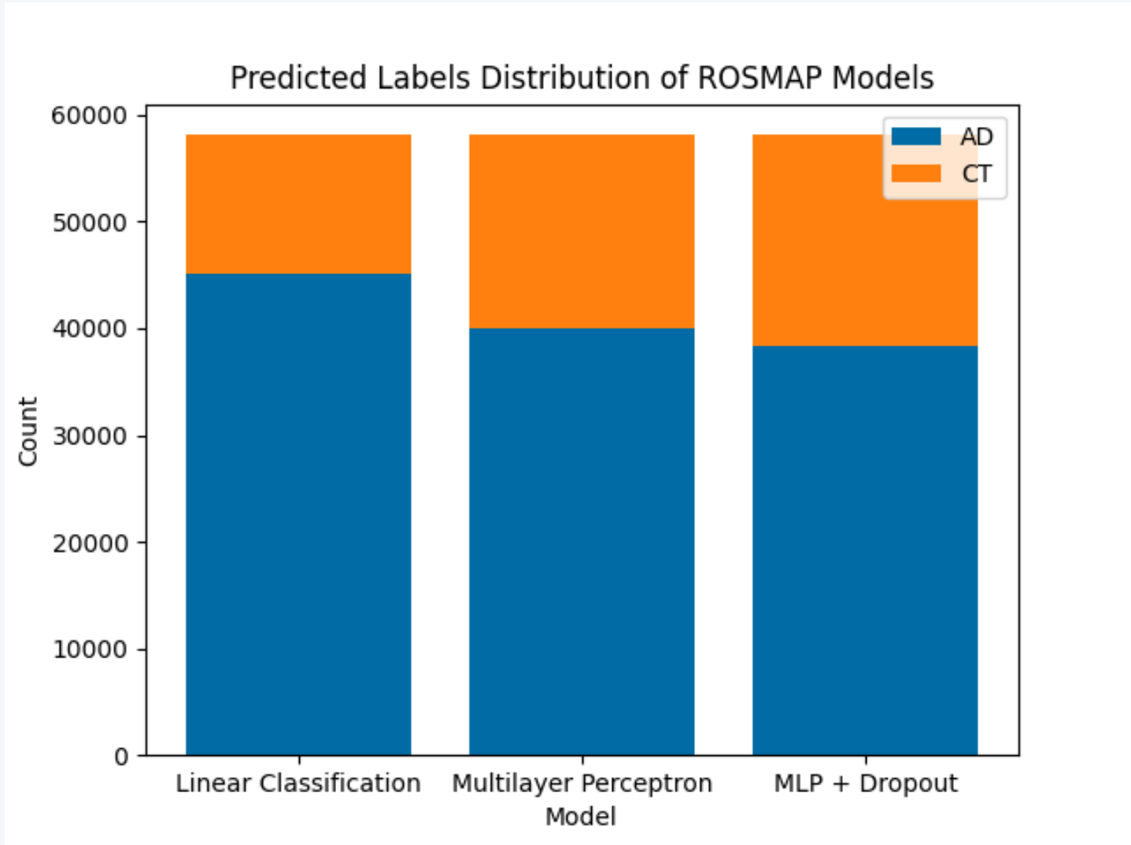

Figure 3. Distribution of predicted labels by ROSMAP

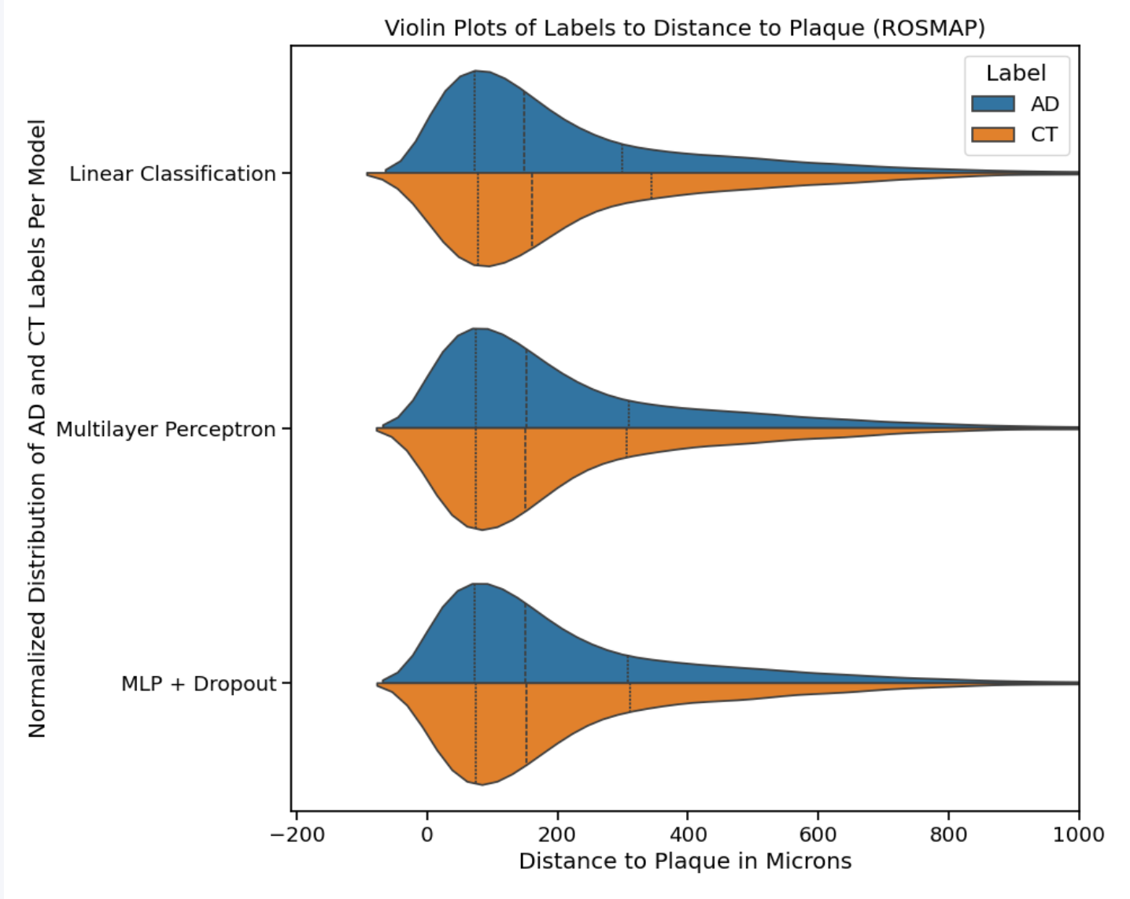

Figure 4. Distance to plaque distribution by labels ROSMAP

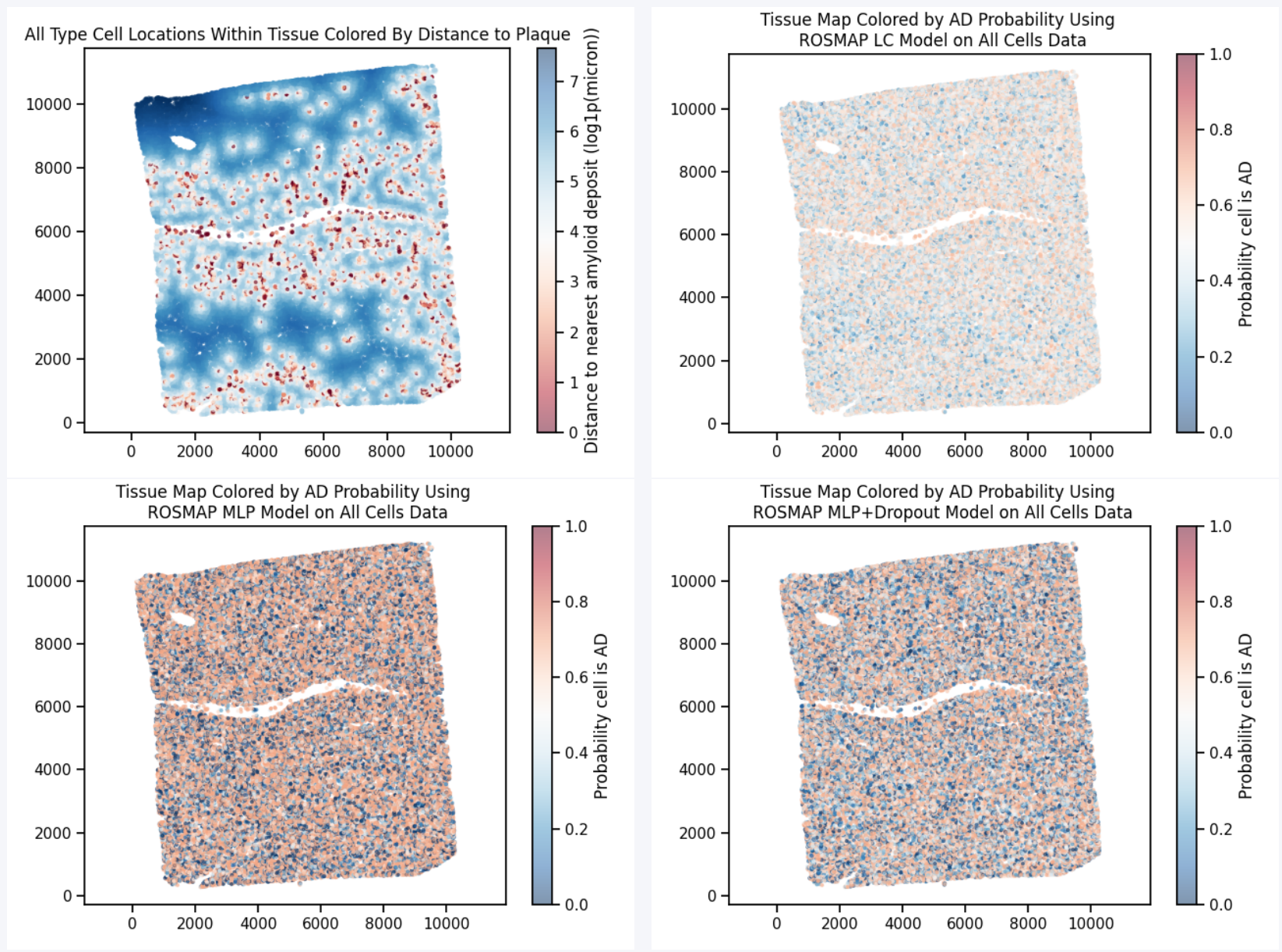

Figure 5. Tissue maps colored by distance to plaque and AD probability from ROSMAP Models

4c. SEA-AD Results

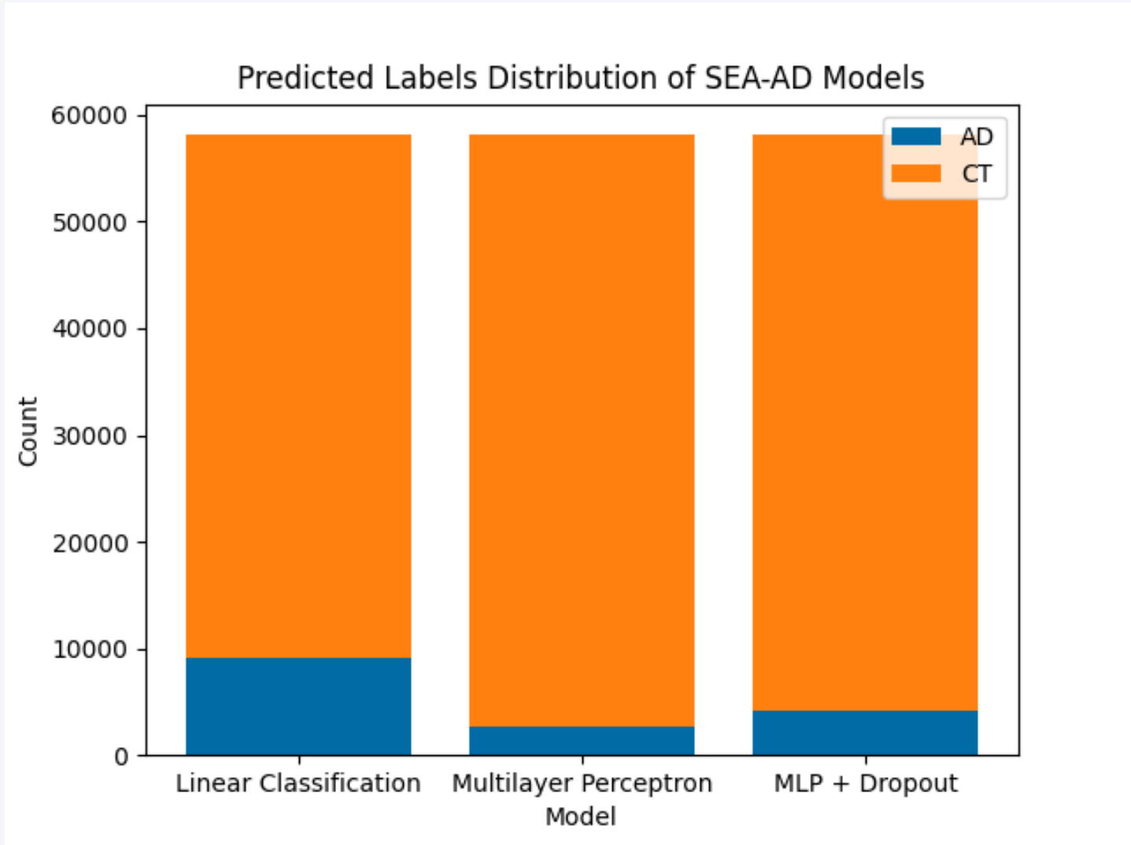

Figure 6. Distribution of predicted labels by SEA-AD

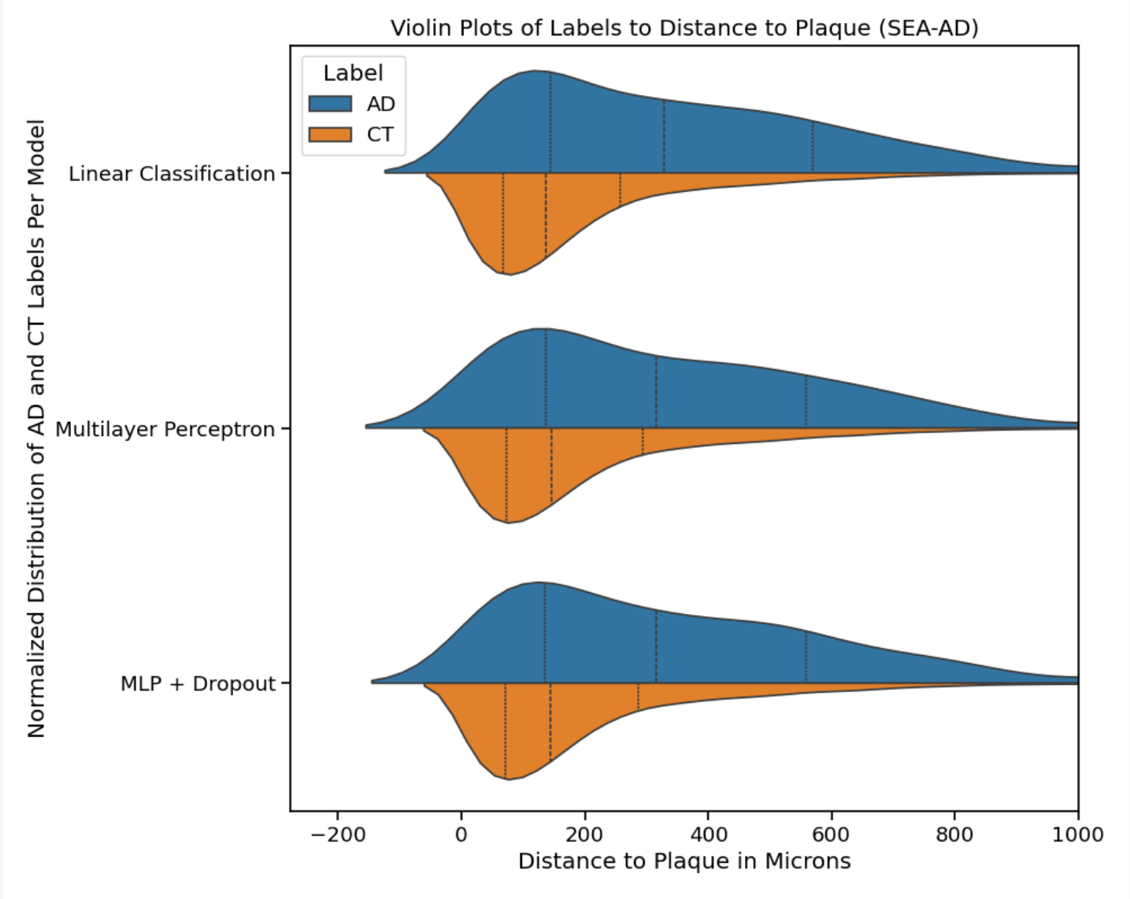

Figure 7. Distance to plaque distribution by labels SEA-AD

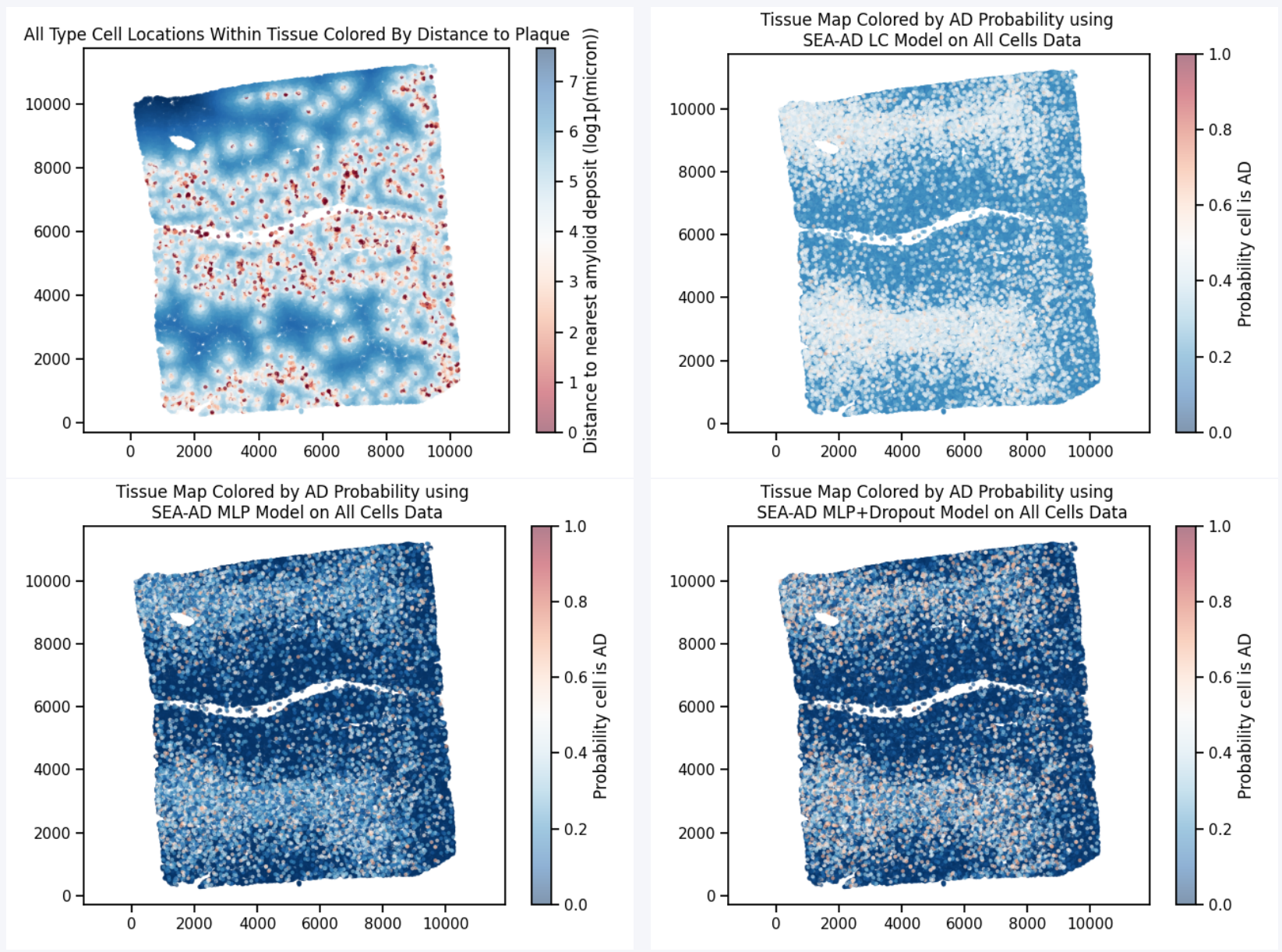

Figure 8. Tissue maps colored by distance to plaque and AD probability from SEA-AD Models

5. Discussion

- High training accuracy, low evaluation accuracy.
- Decreasing evaluation accuracy is likely caused by overfitting on training dataset.
- MLP model seems to perform the best.
- According to Figure 2, imputation is decent, but not perfect. However, it does not show whole picture as validation is incomplete.

- Majority AD cells by ROSMAP models is likely due to AD and CT label distribution in the training data being 2:1. LC model distance distribution matches hypothesis. No spatial pattern identified in AD probability maps.
- Mann-Whitney U test p-value (≈ 0.0) supports LC model claim. However, the Spearman test finds almost no correlation (-0.048) between AD probability and distance to plaque.

- CT label prevalence in SEA-AD results also consequence of imbalance. AD label distance distribution suggests model identifies oligodendrocytes instead. A common spatial pattern is visible, but AD probability is too high. Possible bias in imputation – high AD genes expression.
- U test gives all model predictions a p-value of 1.0, coinciding with high but positive correlation (up to 0.20). This favors the inverse hypothesis, which is unlikely according to literature.

6. Conclusions

- Analysis is not conclusive enough to confirm hypothesis.
- As the converse of the hypothesis is implausible according to literature, the SEA-AD results are likely caused by faulty models or imputation.
- Nevertheless, model evaluation results suggest that MLP without dropout is the preferred model for classification.

7. Limitations

- Xenium data consists of a single donor
  This can introduce bias in the results, such as abnormal spatial distribution of cells. Thus, it is not enough to critically assess the hypothesis' validity in general.
- Low computational and time resources
  Forced to limit imputation to just 5 epochs.
- Simple models with few layers and dimensions
  More complex models could generalize better.

8. Future work

- Attempt experiments with significantly higher number of imputation epochs.
- Tackle model overfitting with regularization and hyper-parameter optimization.
- Prepare datasets more carefully, in order to avoid proportional imbalance between labels and cell types.

References

[1] National Institute on Aging. (2023, Apr.) Alzheimer's disease fact sheet. National Institutes of Health. Accessed: 2025-04-30. [Online]. Available: https://nia.nih.gov/health/alzheimers-disease-fact-sheet

[2] T. Verlaan, G. Bouland, A. Mahfouz, and M. Reinders, "scAGG: Sample-level embedding and classification of Alzheimer's disease from single-nucleus data," *bioRxiv*, p. 2025.01.28.635240, Jan. 2025. [Online]. Available: http://biorxiv.org/content/early/2025/01/30/2025.01.28.635240.abstract

[3] W.-T. Chen, A. Lu, K. Craessaerts, B. Pavie, C. Sala Frigerio, N. Corthout, X. Qian, J. Laláková, M. Kühnemund, I. Voytyuk, L. Wolfs, R. Mancuso, E. Salta, S. Balusu, A. Snellinx, S. Munck, A. Jurek, J. Fernandez Navarro, T. C. Saido, I. Huitinga, J. Lundeberg, M. Fiers, and B. De Strooper, "Spatial transcriptomics and in situ sequencing to study alzheimer's disease," vol. 182, no. 4, pp. 976–991.e19. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867420308151

[4] T. Abdelaal, S. Mourragui, A. Mahfouz, and M. J. T. Reinders, "SpaGE: Spatial gene enhancement using scRNA-seq," vol. 48, no. 18, pp. e107–e107. [Online]. Available: https://doi.org/10.1093/nar/gkaa740

[5] T. Biancalani, G. Scalia, L. Buffoni, R. Avasthi, Z. Lu, A. Sanger, N. Tokcan, C. R. Vanderburg, A. Segerstolpe, M. Zhang, I. Avraham-Davidi, S. Vickovic, M. Nitzan, S. Ma, A. Subramanian, M. Lipinski, J. Buenrostro, N. B. Brown, D. Fanelli, X. Zhuang, E. Z. Macosko, and A. Regev, "Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram," vol. 18, no. 11, pp. 1352–1362. [Online]. Available: https://doi.org/10.1038/s41592-021-01264-7

D.Smenovski@student.tudelft.nl