Supervisors

Timo Verlaan
Roy Lardenoije
Gerand Bouland
Marcel Reinders

CLASSIFYING THE SEVERITY OF ALZHEIMER'S DISEASE

Willem Dieleman
<W.j.j.Dieleman@student.tudelft.nl>

TU Delft

What is Alzheimer's Disease (AD)?

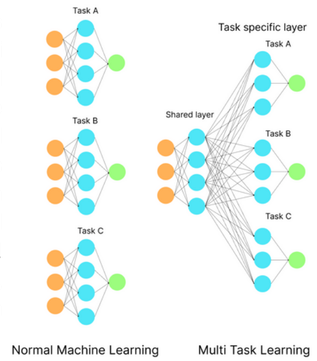
Alzheimer's disease is a complex brain disease and the leading cause of dementia around the world. Treatment options are limited and the true origin is not known. More research at the cellular level is needed to understand the underlying mechanisms.

Single-cell Gene Expression Data

Single-cell gene expression data (scRNA-seq) provides this cellular level data. Research has shown machine learning techniques can use it effectively for things like cancer classification. AD classification has also been done, but this has only been binary, and AD has multiple severities and subtypes

Multi Task Learning

This additional complexity calls for more robust models. **Multi Task Learning** (MTL) could provide this. Its an architecture where multiple things are learned at the same time. If these tasks share commonalities, they can help each other be learned better. It shows a lot of potential for scRNA-seq and AD separately, but it has not been combined with both yet.

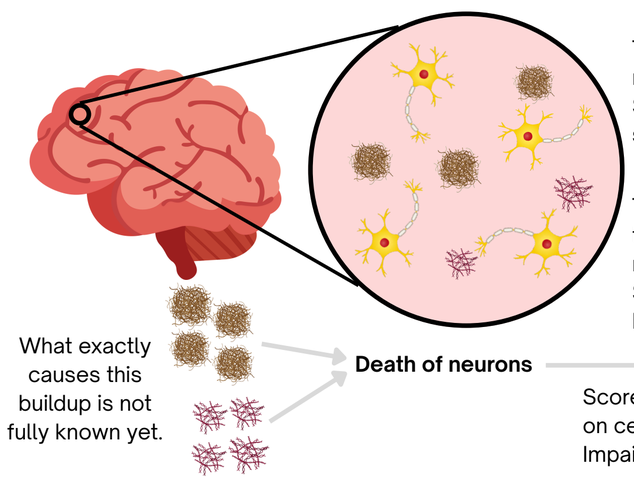


Research question

What (combination of) clinical or pathological measures of AD severity are most accurately classifiable from gene expression data?

AD severity

The development of AD throughout the brain are classified by a buildup of 2 proteins:



Amyloid plaques

These proteins interfere with neuron-to-neuron communication. Severity is denoted by the **CERAD score** (4 classes)

Tau tangles

These proteins interfere with the transportation of molecules to neurons.
Severity and spread is denoted by the **Braak stage** (7 stages)

Death of neurons → **Cogntive Ability**

Score based on the ability to perform on certain tests. 6 classes of Cognitive Impairment exist. Denoted as **Cogdx**.

Methodology

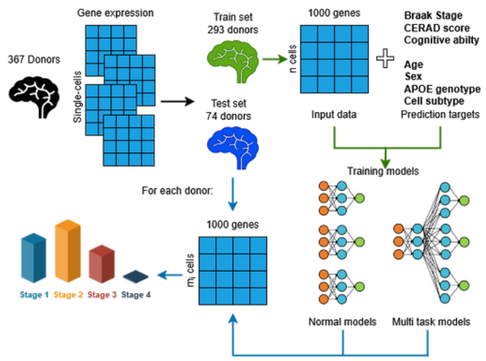
ROSMAP dataset is used, it contains gene expression for 1.6 million brain cells of 450 individuals. After preprocessing **1M cells for 367 donors** remain. Additional metadata linked to AD was selected for MTL: **Age, Sex, APOE genotype, Cell subtype.**

Cell to individual translation

Input data is cells, but things we try to classify are individual level. Some translation is needed for the predictions: Take all cells of an individual, put them through the model, save what the model predicts for each cell, the predicted class with the highest share is the prediction for the donor

Machine learning model

A 3 layer neural-network is used. 1000 best genes are selected. Scores are based on runs with 5-fold cross validation. STL is run once, MTL run twice.



Results

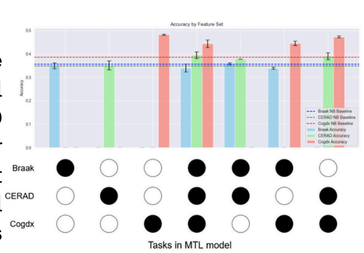
This turns out to be a very **hard task**, Braak and CERAD perform at Naive Bayes level, while Cogdx is slightly above it:

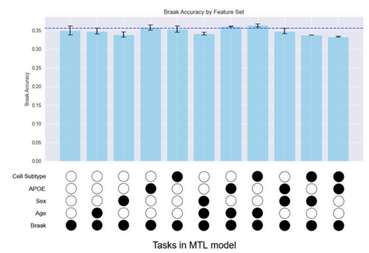
	Microglia	Astrocytes	Cux2+	Cux2-	Inhibitory	Naive Bayes
Braak Stage	0.3501	0.3813	0.3487	0.3123	0.3215	**0.356**
CERAD score	0.3513	0.3761	0.3327	0.3732	0.3925	**0.348**
Cognitive Ability	0.4822	0.4494	0.4693	0.5013	0.4033	**0.386**

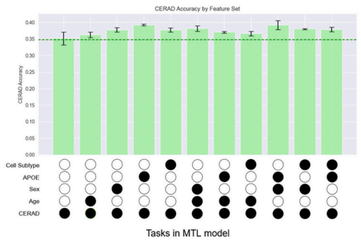
Difference between Braak and CERAD is **not** statistically significant (P = 0.17), same applies to differences between CERAD and Cogdx (P = 0.09). Difference between Braak and Cogdx is **statistically significant** (P = 0.015).

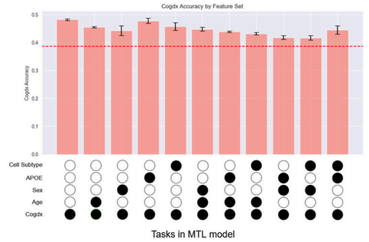
MTL models

MTL appears to have little effect on Braak, a small positive effect for CERAD and a negative effect for Cogdx. Running the best CERAD model 4 additional times vs non MTL gives us no statistical improvement (P = 0.64).









Blue is Braak stage, Green is CERAD score, Red is cogdx, dotted lines indicate Naive Bayes

Cell subtype

We can also analyse performance per cell subtype. Mic.12 and 13 according the literature are connected to the AD severity measures. Same connection can slightly be noticed in the table below. Mic.11 and 15 are outliers due to half of the cells coming from a single donor.

We can use the literature to select only the relevant subtypes for the tasks. Doing this seems to give is an improvement, but Naive Bayes scores also change which explains the entire improvement.

BRAAK	Total	Correct individual		Correct Cellular	
Mic.1	936	309	0,330128	276	0,294872
Mic.2	13269	4592	0,34607	4422	0,333258
Mic.3	7864	2888	0,367243	2654	0,337487
Mic.4	4149	1499	0,361292	1353	0,326103
Mic.5	7923	2655	0,3351	2643	0,333586
Mic.6	6626	1908	0,287957	2095	0,316179
Mic.7	11280	2892	0,256383	2982	0,264362
Mic.8	4424	1451	0,327984	1308	0,29566
Mic.9	3321	1089	0,327913	1132	0,340861
Mic.10	2669	638	0,239041	750	0,281004
Mic.11	564	33	0,058511	47	0,083333
Mic.12	3352	1062	0,316826	1054	0,314439
Mic.13	1966	500	0,254323	646	0,328586
Mic.14	442	204	0,461538	151	0,341629
Mic.15	932	237	0,254292	225	0,241416
Mic.16	729	219	0,300412	233	0,319616
Macrophages	1655	504	0,304532	489	0,295468
Monocytes	678	203	0,29941	245	0,361357

CERAD	Total	Correct individual		Correct Cellular	
Mic.1	936	402	0,429487	286	0,305556
Mic.2	13269	5538	0,417364	4377	0,329867
Mic.3	7864	3728	0,474059	2821	0,358723
Mic.4	4149	1617	0,389732	1361	0,328031
Mic.5	7923	3299	0,416383	2649	0,334343
Mic.6	6626	2356	0,355569	1981	0,298974
Mic.7	11280	4004	0,354965	3457	0,306472
Mic.8	4424	1621	0,36641	1429	0,323011
Mic.9	3321	1223	0,368263	987	0,2972
Mic.10	2669	920	0,344698	862	0,322967
Mic.11	564	394	0,698582	366	0,648936
Mic.12	3352	1599	0,477029	1325	0,395286
Mic.13	1966	1008	0,512716	866	0,440488
Mic.14	442	216	0,488688	122	0,276018
Mic.15	932	537	0,57618	424	0,454936
Mic.16	729	350	0,48011	283	0,388203
Macrophages	1655	691	0,417523	580	0,350453
Monocytes	678	278	0,410029	232	0,342183

Cogdx	Total	Correct individual		Correct Cellular	
Mic.1	936	463	0,494658	374	0,399573
Mic.2	13269	5616	0,423242	5004	0,37712
Mic.3	7864	3497	0,444685	2987	0,379832
Mic.4	4149	2114	0,50952	1677	0,404194
Mic.5	7923	3403	0,429509	3020	0,381169
Mic.6	6626	2807	0,423634	2552	0,385149
Mic.7	11280	4422	0,392021	4241	0,375975
Mic.8	4424	2083	0,470841	1818	0,410994
Mic.9	3321	1394	0,419753	1173	0,353207
Mic.10	2669	1082	0,405395	1000	0,374672
Mic.11	564	30	0,053191	68	0,120567
Mic.12	3352	1504	0,448687	1290	0,384845
Mic.13	1966	880	0,447609	840	0,427263
Mic.14	442	100	0,226244	94	0,21267
Mic.15	932	518	0,555794	400	0,429185
Mic.16	729	400	0,548697	359	0,492455
Macrophages	1655	724	0,437462	623	0,376435
Monocytes	678	295	0,435103	272	0,40118

Correct individual gives all cells for each individual classified correctly, correct cellular is for each cell separately classifying the correct score

	All subtypes	Selected subtypes
Braak Stage	0.3501 +/- 0.0123	0.3561 +/- 0.0076
CERAD Score	0.3515 +/- 0.0191	0.3664 +/- 0.0116
Cognitive Ability	0.4822 +/- 0.0028	0.5214 +/- 0.0083

Discussion

Models perform at a very low accuracy. This is likely because all cells are taken into account, but **not all cells are affected** by the pathology, leading to a lot of noise.
Additionally, cells of only **one region of the brain** are used, which is bad for especially the Braak score.
Due to this, models also suffered a lot from **overfitting** issues.
Additional data like **spatial data** could help filtering out relevant cells.

Conclusions

From this we conclude that Cogdx can be predicted most accurately, but overal performance is low.

MTL does not appear to have a measurable significant positive effect on predictions with the current architeture.

Cell type analysis shows potential, but needs to be explored further.