A Multi Task Learning approach to classifying the severity of Alzheimer's disease using single-cell gene expression data

Willem Dieleman[1]
Supervisor(s): Marcel Reinders[1]**, Timo Verlaan**[1]**, Roy Lardenoije**[1]**,**
Gerard Bouland[1]
[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 27, 2025

Name of the student: Willem Dieleman
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Timo Verlaan, Roy Lardenoije, Gerard Bouland, Ricardo Marroquim

Abstract

Alzheimer's Disease (AD) is a complex heterogeneous disease and is the leading cause of dementia around the world. Treatment options remain limited and the underlying mechanisms are not yet fully understood. To get more insight on this celular level, single-cell gene expression data can be used. It has proven to be effective with machine learning for tasks like cell type classification. While prior studies have explored AD classification using scRNA-seq, this has only been a binary classification. Severity of AD is classified using multiple measures, ranging from cognitive ability scores, to neuro pathological measures. This research explores the possibility of expanding the binary prediction of AD by including these measures for AD severity. In addition, given that these measures are associated, we also investigate if Multi Task Learning (MTL) models can improve the predictions by learning multiple AD related data points. If successful, this approach can give additional analysis into key tasks, genes and/or cells (sub)types that drive the models, which would lead to more possibilities for personalized treatment options, alongside more insight into the development of AD in the brain. We used a three-layer neural network architecture alongside a translation from cellular level to individual level to make individual-level predictions. Results show that Cognitive Ability can be classified best, but overal performance is only slightly above Naive Bayes. Furthermore, MTL does not appear to have any measurable positive effect on scores compared to single task models. A link to the github repository is available at https://github.com/WillemDieleman/ADseverityCSE3000.

1 Introduction

Alzheimer's disease (AD) is a complex neurodegenerative disease and is the leading cause of dementia [1]. No cure for AD exists and the treatment options are very limited. While the exact cause of AD is not fully understood, it is strongly associated with an excess amount of amyloid plaques and tau tangles in the brain, contribution the death of neurons [1]. The reason behind this behind this buildup remains unclear [2], though research suggests factors like genetics and lifestyle contribute to the risk of developing AD [3].

Additional research is needed at the cellular level to fully understand the underlying mechanisms behind AD and how it develops throughout the brain [1]. One way of doing this is by getting single-cell gene expression data (scRNA-seq) from individuals with and without AD. Unlike bulk RNA-seq, which averages gene expression for all cells of an individual, scRNA-seq gives us that cellular level needed for AD research, where patterns of affection between different cell (sub)types can be identified. To achieve this, this data can be used to train machine learning models to classify AD-related characteristics. Similar research into cancer cells has shown that this is possible. A research used scRNA-seq data to classify cancer cell types and achieved accuracies of up to 99% . They found neural networks had the best performance for binary and multi-class classification [4]. PanClassif, a similar research into cancer cell classification, found similar results, with k-Nearest Neighbors, Random Forest and Neural Networks performing best at around 99% accuracy for both binary and multi-class classification [5].

Cancers, however, are quite different compared to AD. While cancers can also be heterogeneous, especially in later stages, they often focus around tumor cell populations [6]. AD lacks these discrete populations, and there are no clear 'AD cells' as opposed to tumor cells. AD affects multiple regions and cell subtypes in the brain [1], with especially 2 subtypes of microglia cells being strongly associated with AD [7]. To classify the severity of AD, multiple pathological or clinical measures are used. The main three measures used in this research are the Braak Stage, indicating the spread and severity of tau tangles in the brain; the CERAD Score, indicating the abundance of amyloid plaques; and the Cognitive Ability of a donor. This additional complexity in severity requires more robust models that are capable of highlighting the difference in these measures.

Multi-Task Learning (MTL) is a machine learning approach in which a model learns to predict outputs at the same time. MTL works best with tasks that use the same input data and share commonalities [8]. In the context of AD, this can prove to be useful, as the input data is the same, namely the scRNA-seq data, and the tasks are all related to AD. MTL uses this association across tasks to potentially improving performance. A study using MTL and scRNA-seq has shown that it can be used effectively. They used MTL models to predict cell types along side learning different subpopulations. This led to an increase in performance compared to other state of the art models, while also reducing computing time [9]. Figure 1 provides a visual comparison between MTL and conventional machine learning.

MTL has also been used in combination with AD classification, but these have mainly been focused on MRI-scan input data, with a study from Zhang et al. achieving a 4% accuracy improvement in classifying AD and healthy patients [10]. While promising, these approaches do not include the cellular mechanics that is needed for identifying the underlying mechanisms of AD and are limited to healthy vs AD classification.

While MTL is not widely used in combination with scRNA-seq and AD, a related study, scAGG, whose premise is very similar to our research, used the same scRNA-seq data as our research to create a sample-level embedding to classify AD. They achieved an accuracy of around 75% classifying healthy and AD donors [11]. Our research aims to extend this by not only looking at more donors, but also finding out which clinical or pathological measure that quantifies AD can be classified best with the addition of using MTL to look at combinations of measures. This will be explored in 2 parts:

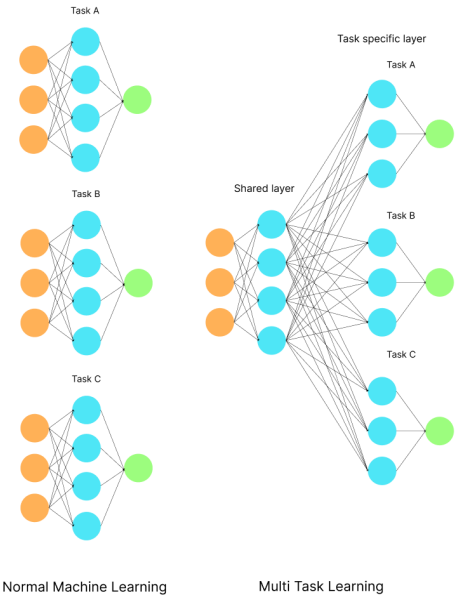

Figure 1: Diagram of neural networks and multi task neural networks

1. Single-task - where we create and train a model for each of the measures separately and evaluate which of these measures can be predicted best.

2. Multi-task - where we apply MTL principles to enhance these basic models to work with additional measures, and finding out if certain (combinations of) tasks improve the performance.

The ROSMAP dataset is used in this research. It contains the scRNA-seq data for 465 individuals, includes the pathological and clinical measures required for AD severity and contains additional metadata of donors, like age and sex, that have a connection to AD [12].

From this data, we propose the following hypothesis:

1. Since the input data contains gene expression data from brain tissue, we hypothesize that measures linked to processes inside the brain can be classified more accurately than measures outside of the brain.

2. Additionally, given that the amyloid cascading typically goes from plaques $\rightarrow$ tangles $\rightarrow$ cognitive decline [7], we expect MTL to have a measurable improvement on the Braak Stage and Cognitive Ability if trained in combination with the CERAD Score or other measures. We expect no measurable improvement in performance using MTL on the CERAD Score.

To test these hypotheses, we train these models using 5-fold cross validation on multiple runs and use t-tests to test for statistical significance.

2 Materials & Methodology

2.1 ROSMAP Dataset

The main dataset used in our research is the single-nucleus gene expression data from the ROSMAP dataset. Data containing 1.6 million cells from 465 donors is available [12]. Furthermore, for each donor additional metadata is available, these include the 3 AD severity measures mentioned in the introduction:

Braak Staging which indicated the abundance and spread of the tau tangles throughout the brain. Tau tangles handle the transfer of nutrients and more to the neurons and is a factor inside the neurons [2]. Braak staging ranges from 0 to 6, where 0 is no tangles and 6 is tangles throughout all parts of the brain [13] [14].

CERAD Score which indicates the abundance of the amyloid plaques in the brain. These plaques interfere in neuron-to-neuron communication and is a factor outside of the neurons. 4 scores exist: Definitive AD, Probable AD, Possible AD and No AD [15] [16].

Cognitive Diagnosis (Cogdx) which indicates the ability a patient is able to function based on cognitive tests alongside diagnoses from neuropsychologists and clinicians. Individuals are divided into 6 classes, which can be seen in table 1 [15].

Value	Coding
1	NCI: No cognitive impairment
2	MCI: Mild cognitive impairment, no other condition contributing to CI
3	MCI+: Mild cognitive impairment AND another condition contributing to CI
4	AD: Alzheimer's dementia, no other condition contributing to CI
5	AD+: Alzheimer's dementia AND other condition contributing to CI
6	Other dementia: Other primary cause of dementia

Table 1: Cognitive Diagnosis class division

Additional information linked to AD is also available in the dataset. This data will be used as additional tasks for the MTL part. The 4 characteristics we picked are:

Sex Females are more likely to develop AD [17].

Age AD is an age-related disease [18].

APOE Genotype A known genetic risk factor of AD [19]

Cell Subtype Certain subtypes are associated with the AD measures [7].

To compare our models against existing binary AD classification, the definition of AD from Wang et al. (2021) [20] is used. They filter out the most extreme AD and control cases with the following definition:

$$\begin{cases} Control & \text{if cogdx = 1, braaksc} \leq 3\text{, ceradsc} \geq 3 \\ AD & \text{if cogdx = 4, braaksc} \geq 4\text{, ceradsc} \leq 2 \\ OTHER & \text{else} \end{cases}$$

2.2 Preprocessing

Following the AD guidelines of Wang et al. (2022) [21], doublets and low-quality genes and cells were removed, alongside selecting the 4000 most variable genes [7]. Additionally, donors with a PMI of more than 12 hours are removed

[22]. Individuals with missing metadata are also removed. Due to limited samples for Braak stages 0, 1, and 6, they were merged into a 'Low stage' (Braak stages 0, 1, 2) and a 'High stage' (Braak stages 5, 6). The same applied for Cogdx, with limited samples for values 3 and 5. They were merged into a group for 'Mild Cognitive Impairment' (values 2, 3) and 'Alzheimer's Disease' (values 4, 5). Cogdx value 6 was excluded. The age metadata contains '90+' values for any individual older than 90 years. These were replaced with the value of 90, and then the ages were mean-normalized. Oligodendroglia cells were excluded. This leaves us with 367 individuals totaling 994,827 cells. All Multi Task models are trained are only trained on microglia cells, of which we have 72,779 cells. Single-Task models are trained on all celltypes; They include microglia, astrocytes, inhibitory neurons, and excitatory neurons (cux2+, cux2-).

2.3 Task representations

In our reseach, we used seven separate prediction tasks are used in the MTL models. These are:

1. **Braak Stage** An ordinal classification task with 4 classes: Low stage, 3, 4, High stage.
2. **CERAD score** An ordinal classification task with 4 classes: Definitive AD, Probable AD, Possible AD, No AD.
3. **Cogdx** An ordinal classification task with 3 classes: No CI, Mild CI, AD.
4. **Age** A regression task with a continuous value.
5. **Sex** A binary classification task: Female (F) and Male (M).
6. **APOE genotype** A classification task with 6 classes: 22, 23, 24, 33, 34, 44.
7. **Cell subtype** A classification task. For microglia, cells consists of 18 sub types: Mic.1 - 16, Monocytes, Macrophages.

2.4 Feature Selection

To reduce the dimensionality of the data further, an Analysis of Variance (ANOVA) test was used for feature selection. ANOVA is a statistical method that tries to identify features that have a large variance across the target classes. This is similar to what was used in PanClassif [5]. Using this method and to avoid data leakage between the test and train sets, the 1000 best features of the training set were calculated, and then the same features were selected for the train set.

For feature selection of the MTL part, a similar approach to Kim et al. (2019) [4] was used. For each task, the ANOVA test was used to determine the 1000 best features. These features are compiled in a list. Once this has been calculated for all tasks, the 1000 most picked features in the list are selected for the final feature selection.

2.5 Experimental Setup

We use a three-layer neural network for both the single task learning (STL) and the multi task learning (MTL) models.

Hidden layers Both the STL and MTL models consist of three hidden layers with size {1024, 256, 64}. In the MTL model, the first two layers (1024 and 256 nodes) are shared across tasks, and the third layer (64 nodes) is task specific. All hidden layers used ReLU activation functions and batch normalization.

Loss functions For regression tasks we used Mean Squared Error (MSE) loss. For classification tasks we used Cross Entropy loss. For ordinal classification tasks, we used Binary Cross Entropy With Logits loss.

Optimizer All models used the Adam optimizer with a learning rate of 0.001 and a weight decay of 0.001.

Early stopping Each model was trained for up to 100 epochs, with early stopping using a patience of 3.

Train/test split We used 5-fold stratified cross validation. In the case of MTL, composite labels were created by concatenating the classes per task into a single label used as the stratify target.

Random seed All random seeds used were set to 42.

Cell versus Individual

For each donor, we have the measures and other metadata as well as on average around 2500 cells available. The AD severity measures we are trying to classify are on the individual level, while the input data will be on the cellular level. That means some translation needs to be made from the cellular level to the individual level. To achieve this, the train/test split will be done on individual level; For the training set, all the cells of the individuals are taken separately as an input vector with the target being the selected measure of the individual the cell belongs to. For testing, from each individual in the test set all cells are taken and are separately put through the model. The prediction for each cell is saved, after which a distribution is made per individuals for the counts per predicted class. The class with the highest share of the predictions will be selected as guess for the individual.

Evaluation

Models are evaluated based on the individual level. This makes use of the translation explained in the previous paragraph; For each individual in the test set a prediction is made, which is compared against the true answer. We used the accuracy metric to compare models.

2.6 Implementation details

Scanpy and Pandas was used to process the input files. Pytorch and SKlearn was used to create, train and test the models. Matplotlib, Seaborn, Numpy and SHAP were used to process, analyze and plot the data. All results that used microglia data was run locally on a system with 8 cores and 32 GB or RAM running Windows 11. Any data involving other cell type data was run on the Delft AI Cluster (DAIC) [23].

3 Results

We used the single-nucleus RNA-seq dataset from ROSMAP [12]. After preprocessing the data, we are left with 376 donors totaling to 994,827 cells. The 1000 best genes are selected for the tasks (see Methodology). Models are trained on

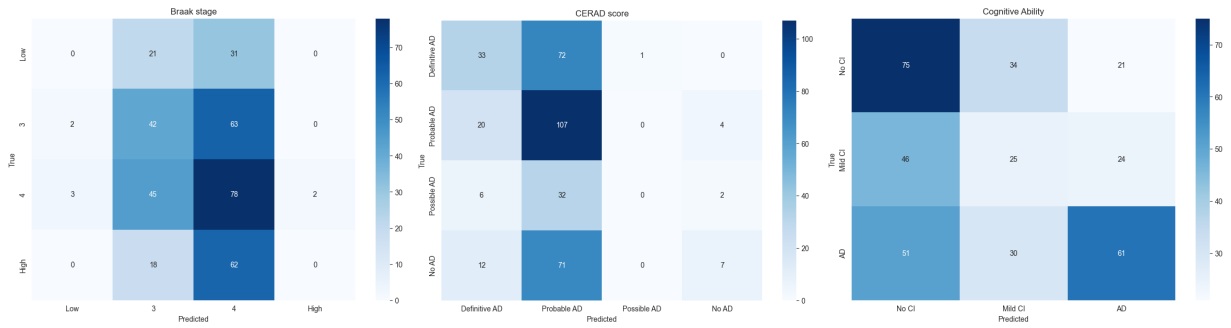

Figure 2: Confusion matrices for Braak stage, CERAD score and Cogdx. Data is based on a single run using 5-fold cross validation using microglia data.

cellular level, and then tested and compared using accuracy on individual level (see Methodology). Single Task models are evaluated on multiple cell types over a single 5-fold cross validation run. Multi Task models are only evaluated on microglia cells over two 5-fold cross validation runs, after which scores are averaged.

3.1 Single Task Models

To compare the accuracy between different measures, the Naive Bayes baseline is needed. For Braak Staging, the most common stage, which is stage 4, contains 35.6% of the individuals. For CERAD score, the most common class is 'Probable AD', with 34.8% of individuals and for Cognitive Ability its 'Alzheimer's Disease' with 38.6%.

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

Table 2: Accuracies for the AD severity measures based on the cell type as input. Results are based on a single run using stratified 5-fold cross validation.

From the results in table 2, Braak Stage classification achieves an average accuracy of 34.3 $\pm$ 2.4% over all cell types. This is within margin of error for the Naive Bayes baseline. Looking at a confusion matrix for the Braak stage in figure 2, it can be seen that it only focuses its guesses on the 2 most common stages, stages 3 and 4. What can also be observed is that the majority of the 'high stages' are at least classified at stage 4, but opposite is not true for the 'low stages'.

For the CERAD score, the average accuracy in 2 is 36.5 $\pm$ 2.1% over all cell types. This is also within the margin of error of the Naive Bayes baseline. A similar story as the Braak Stage can be seen in the confusion matrix in figure 2, with predictions being classified around the 2 most common classes. Noteworthy are the few No AD cases being correctly classified, showing potential of this task.

For Cognitive Ability, the average accuracy in table 2 is 46.1 $\pm$ 3.3%. This is significantly higher than the Naive Bayes baseline. Looking at the confusion matrix in figure 2 a clear distinction for AD predictions can be seen. The same to a smaller extend applies to the No CI class, while for the Mild CI class, it seems to guess randomly.

After normalizing the results with the Naive Bayes baselines, we ran a paired t-test for statistical significance. CERAD and Braak show no statistical significance (P = 0.17) difference in performance. The same applies to CERAD and Cogdx (P = 0.09). Braak and Cogdx do show statistical significance (P = 0.015). From this we can reject the first hypothesis, as the measures linked to processes inside the brain cannot be classified more accurately than those outside of the brain.

We also compared our model with scAGG [11] to classify the most extreme healthy and AD cases. Using this method left us with 142 individuals (51 Control, 91 AD). Our results are only based on microglia cells, while scAGG uses all cell types, so the comparison is not perfect. ScAGG achieves a peak accuracy of around 73 $\pm$ 7%. Our model achieves an accuracy of 75.7 $\pm$ 2.2% over 20 runs using 5-fold cross validation.

3.2 Multi Task Models

All MTL models are only trained and tested on the microglia cell data. All combinations up to 2 additional measures are tested. Models were trained for 2 runs using 5-fold cross validations. Accuracies are averaged over the 2 runs.

From the data in figure 3, MTL does not appear to have any effect on the accuracy of the Braak stage predictions. All scores fall within the margin of error of the single task model and the Naive Bayes baseline.

For CERAD score, an improvement of using MTL can be noticed. Combinations like CERAD + Braak + Cogdx, CERAD + Apoe and CERAD + Sex + APOE performed the best with an accuracy around 39%, which around 4 points higher than the Naive Bayes baseline, but looking at the high variability of the single task model, this could be due to variance.

For Cogdx, MTL seems to reduce performance, as all combinations perform worse than the single task model, with the worst combination of Cogdx + Sex + Cell subtype performing 7 points lower than the baseline. All combinations still perform better than Naive Bayes.

Looking at the SHAP values of the tasks in figure 4, which indicate if the presence of certain tasks (red is included, blue

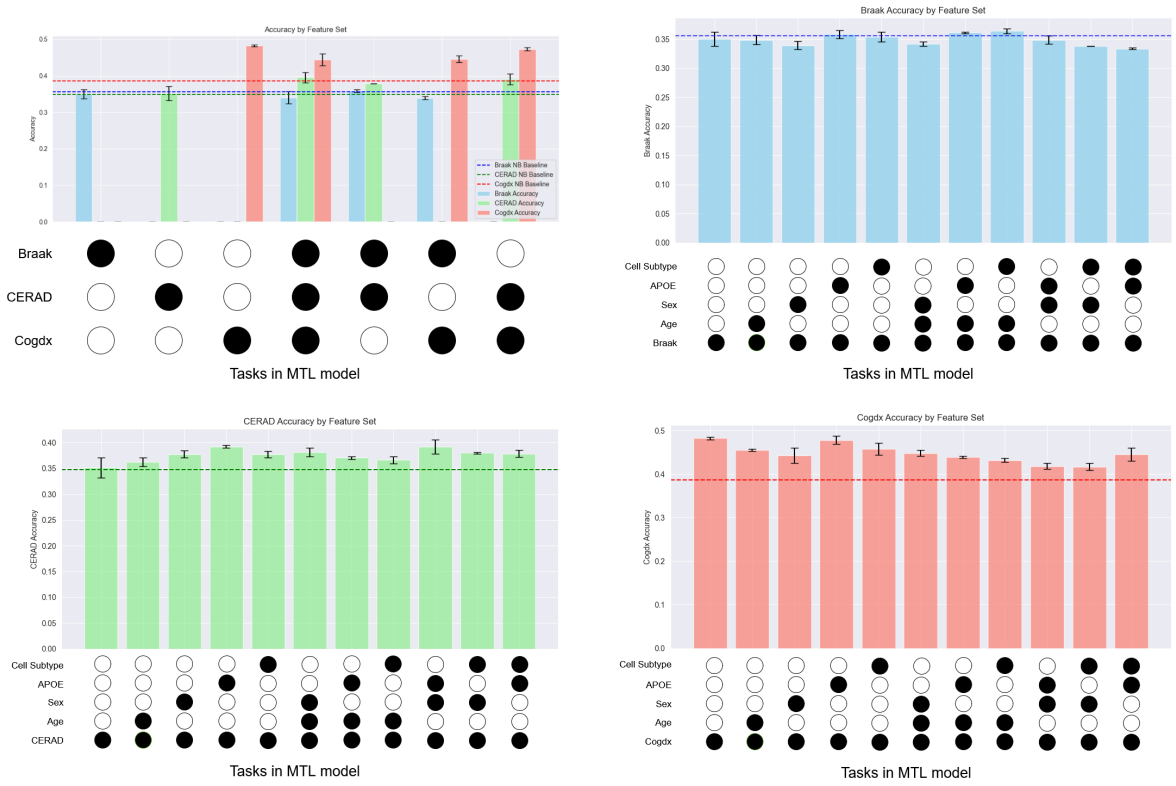

Figure 3: Accuracies for Braak stage (Blue), CERAD score (Green) and Cogdx (Red) per MTL taskset. Data is based on the average of 2 runs using 5-fold cross validation using microglia data. Dotted lines indicate Naive Bayes baselines.

is not included) has an effect on prediction performance. For Braak, all tasks seem to have a very small effect on the model, with the only relevant result being that sex consistently reduces performance. For CERAD, it can be seen that both the inclusion of APOE and sex tasks consistently improve performance. For Cogdx, we see that the inclusion of any additional tasks reduces performance, with sex reducing performance the most.

Running the best combination for the CERAD score, CERAD + Sex + APOE 4 additional times, and comparing it to 4 additional runs of the single task model shows no statistical significance (P = 0.64) on a two-sided t-test confirming that these results need to be run multiple times to better compare performance.

The additional tasks included in the MTL model also had their performance measured. Sex had an average accuracy of 99%, APOE genotype had an average accuracy of 57.4% which is very close to the Naive Bayes baseline for APOE genotype of 57.5%. Cell subtype had an average accuracy of 82%. Age had an MSE loss on average of around 1.2, with an age std being 4.5, meaning it on average was 6-7 years away from the true answer, but this is likely biased since 42.5% of individuals had their age listed as '90+' and were mapped to 90.

Looking at the correlation matrix in figure 5, the 3 AD severity measures are strongly correlated, and age and APOE being slightly less connected to the measures, while sex is the least correlated. MTL typically works well on tasks that

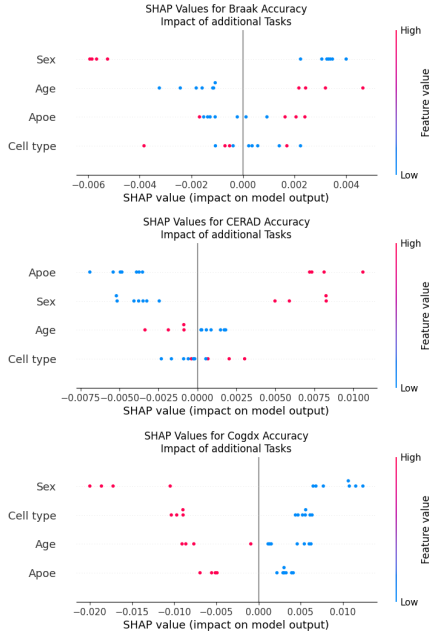

Figure 4: SHAP values calculated for all additional tasks for each of the AD severity measures. Red means included in the task list, while blue means not included. Data is based on 2 runs using 5-fold cross validation on microglia data.

share commonalities [8], while it has also shown to have an improvement in performance on unrelated tasks if they are trained in the same input data [24]. In theory this should mean that MTL would help in predictions of the measures, but we found that this improvement cannot be statistically confirmed based on the limited runs we have done. We believe this has to deal with the low performance of the model in general. Due to the limited runs, we cannot conclusively reject the second hypothesis, but these results suggest it its current model architecture, MTL does not improve predictions.

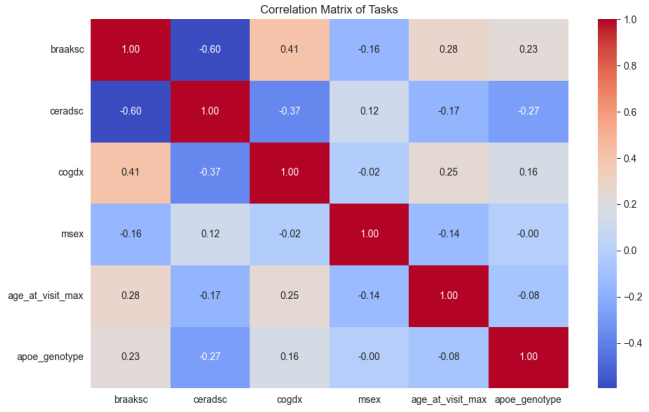

Figure 5: Correlation matrix for all tasks except cell subtype based on the preprocessing outlined in the methodology. Data is based on 367 individuals.

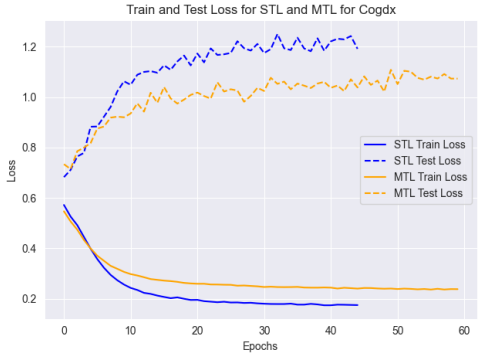

Figure 6: Losses for Cogdx task per epoch. The MTL model is one with tasks [Cogdx, CERAD]. Second image is when Data is based on the first fold using microglia data.

Looking into why these models performed at a very low level, we plotted the train and test loss of Cogdx for both a Single Task model, as well as a Multi Task model in combination with Cogdx and CERAD. This data can be seen in figure 6. This shows clear signs of overfitting. What does appear is the difference in loss between STL and MTL. The MTL test loss is lower than that of STL, but looking back in the results in figure 3, this does not translate into an improvement in performance. Even limiting training to 50 batches of 64 cells per training step did not seem to help, as can be seen in figure 7.

Figure 7: Training and test loss per epoch for Cogdx STL model with a limit of 50 batches per training step. Data is based on the first fold using microglia data.

3.3 Cell types

Additional analysis can be made using the microglia subtype of all the cells. Microglia data consists of 16 subtypes, alongside monocytes and macrophages [7]. A Multi Task model consisting of the tasks [Braak stage, CERAD score, Cogdx, Cell subtype] was trained, and at the end of every fold all individuals in the test set that were classified correctly had their cell subtypes was saved. Additionally all cells in the train set were individually checked if they classified the correct score. The data for this can be found in table 3. The data from all 5 folds is combined.

According to a study from Green et al. (2024) [7], microglia subtypes Mic.12 and Mic.13 have strong associations with all 3 measures, with Mic.12 having a strong association with age, while Mic.13 had an association with amyloid plaques and tau tangles, and though association cognitive decline. From our results, that association can be slightly noticed, with both of the subtypes doing above average on the cellular level for all the measures. The individual level for Braak appears to be a bit low, but that can be explained by the low performance of the model. From our data it can also be noticed that Mic.11 has some unusual behavior, being the best sub type for CERAD score and the worst for both Braak and Cogdx. Mic.11 is associated with stress responses, and only has a small connection to plaques and tangles [7]. A single donor contains almost half (228) of all the Mic.11 cells, meaning this result is likely an outlier. The same applies to Mic.15, which deals with inflammatory reactions and is associated with Cognitive Ability [7], which would explain the high performance in the Cogdx part, but for this subtype, almost half (454) of the cells also come from a single donor.

Using the data from Green et al. [7], we can make a selection of subtypes that are associated with the 3 main measures. For the Braak stage this selection would consist of the following subtypes: [Mic.3, Mic.4, Mic.5, Mic.11, Mic.12, Mic.13, Mic.15]. For the CERAD score this selection consists of [Mic.2, Mic.12, Mic.13, Mic.14, Mic.15] and for Cogdx the selection is [Mic.11, Mic.12, Mic.13, Mic.15]. The exact preprocessing and model architecture as outlined in the methodology section are used, with the only addition being that individuals with less than 20 cells have been removed. Doing this leaves us with around 26k cells for 328 individuals for Braak, 18k cells and 262 individuals for CERAD, and 5.3k cells for 117 individuals for Cogdx. From this data, the Naive Bayes Baseline slightly changes. For the Braak stage this now

BRAAK	Total	Correct individual		Correct Cellular		CERAD	Total	Correct individual		Correct Cellular		Cogdx	Total	Correct individual		Correct Cellular	
Mic.1	936	309	0,330128	276	0,294872	Mic.1	936	402	0,429487	286	0,305556	Mic.1	936	463	0,494658	374	0,399573
Mic.2	13269	4592	0,34607	4422	0,333258	Mic.2	13269	5538	0,417364	4377	0,329867	Mic.2	13269	5616	0,423242	5004	0,37712
Mic.3	7864	2888	0,367243	2654	0,337487	Mic.3	7864	3728	0,474059	2821	0,358723	Mic.3	7864	3497	0,444685	2987	0,379832
Mic.4	4149	1499	0,361292	1353	0,326103	Mic.4	4149	1617	0,389732	1361	0,328031	Mic.4	4149	2114	0,50952	1677	0,404194
Mic.5	7923	2655	0,3351	2643	0,333586	Mic.5	7923	3299	0,416383	2649	0,334343	Mic.5	7923	3403	0,429509	3020	0,381169
Mic.6	6626	1908	0,287957	2095	0,316179	Mic.6	6626	2356	0,355569	1981	0,298974	Mic.6	6626	2807	0,423634	2552	0,385149
Mic.7	11280	2892	0,256383	2982	0,264362	Mic.7	11280	4004	0,354965	3457	0,306472	Mic.7	11280	4422	0,392021	4241	0,375975
Mic.8	4424	1451	0,327984	1308	0,29566	Mic.8	4424	1621	0,36641	1429	0,323011	Mic.8	4424	2083	0,470841	1818	0,41094
Mic.9	3321	1089	0,327913	1132	0,340861	Mic.9	3321	1223	0,368263	987	0,2972	Mic.9	3321	1394	0,419753	1173	0,353207
Mic.10	2669	638	0,239041	750	0,281004	Mic.10	2669	920	0,344698	862	0,322967	Mic.10	2669	1082	0,405395	1000	0,374672
Mic.11	564	33	0,058511	47	0,083333	Mic.11	564	394	0,698582	366	0,648936	Mic.11	564	30	0,053191	68	0,120567
Mic.12	3352	1062	0,316826	1054	0,314439	Mic.12	3352	1599	0,477029	1325	0,395286	Mic.12	3352	1504	0,448687	1290	0,384845
Mic.13	1966	500	0,254323	646	0,328586	Mic.13	1966	1008	0,512716	866	0,440488	Mic.13	1966	880	0,447609	840	0,427263
Mic.14	442	204	0,461538	151	0,341629	Mic.14	442	216	0,488688	122	0,276018	Mic.14	442	100	0,226244	94	0,21267
Mic.15	932	237	0,254292	225	0,241416	Mic.15	932	537	0,57618	424	0,454936	Mic.15	932	518	0,555794	400	0,429185
Mic.16	729	219	0,300412	233	0,319616	Mic.16	729	350	0,48011	283	0,388203	Mic.16	729	400	0,548697	359	0,492455
Macrophages	1655	504	0,304532	489	0,295468	Macrophages	1655	691	0,417523	580	0,350453	Macrophages	1655	724	0,437462	623	0,376435
Monocytes	678	203	0,29941	245	0,361357	Monocytes	678	278	0,410029	232	0,342183	Monocytes	678	295	0,435103	272	0,40118

Table 3: Accuracies split up per microglia subtype for each of the AD severity measures. Columns outline the total amount of cells per subtype, amount of cells per subtype from all individuals in the test set who are classified correctly, alongside the amount of cells per subtype from all cells in the test set that classified the correct score. Data is based on a Multi Task Model with Braak stage, CERAD score, Cogdx and cell subtype as tasks. Data is from the combination of 5-fold cross validation results.

is 34.5%, CERAD score it is 36.6% and Cogdx its 53.0%. Training these models again for 2 runs gives us the results as outlined in table 4.

	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

Table 4: Accuracies for the main measures for all microglia cells compared to a selection. Data is based on a selection of microglia subtypes and data is averaged over 2 runs using 5-fold cross validation.

No real improvement can be seen for Braak and CERAD with the models still performing similarly to the Naive Bayes baseline. For Cogdx, which used to perform significantly better than the baseline, now performs at the Naive Bayes baseline, meaning it actually lost performance.

4 Discussion

We looked into the possibility of classifying the severity of AD by trying to classify the Braak Stage, CERAD score and Cogntive ability using single-cell gene expression data. We combined this with Multi Task Learning principles to look for improvements in performance. We found that these tasks are very difficult for the model architecture we created, with Braak and CERAD performing close to Naive Bayes level, while Cogntive ability performed slightly above it. MTL appeared to have very limited effects on Braak and CERAD, while reducing performance for Cogntive ability. MTL does seem to help against overfitting, but this has to be explored further to conclusively confirm this.

We believe the bad performance could be due to our translation from cellular to individual level. In our architecture, we hope that the enough cells are affected by the pathology that we can classify the individual correctly, but this is often not the situation. Take for example the Braak Stage; One important factor for determining which stage an individual has, is the spread of tau tangles throughout various parts of the brain. With low stages, they are just the brain stem, while in the highest stage, they are spread around the entire brain [13]. The cells used in this research come from the dorsolateral prefrontal cortex [7], which is only a single region of the brain and only gets the spread of the tangles around Braak stage 5 [13]. This means that for previous stages, the region where we get our data from is not affected by the tau tangles yet, which could explain why the model does not perform better than Naive Bayes. This can be improved by having data from multiple parts of the brain. Furthermore, the Braak stage indicates processes inside the neuron [2]. Using microglia data would naturally have more difficulty with this, as it is not a neuron.

This could also explain why classifying the CERAD score appears to perform ever so slightly better, as it denotes the amyloid plaques, which is a factor outside of the neurons, which the microglia interact with more. Cells closer to the amyloid plaques are affected more [25]. Thus, to improve the CERAD classification, spatial data about the location of each cell and distance to pathology can be used to only select the cells that are close enough to be affected and using those to train the model. Using a model like SpaGE [26], or using the a spatial dataset could enhance predictions, but has not been done in this research.

Finally, the bad translation could explain the issues with overfitting issues we ran into. It labels all the unaffected cells to a certain score in the training, while in the testing, very similar unaffected cells are linked to a different score and thus classified incorrectly. This explains the rise in loss on the test set immediately from the first epoch onward. To try and circumvent this, you can again try the spatial data, or extending the cellular level analysis by finding patterns in cells which show cells are not affected.

Improvement of the model architecture will be critical for improvement in performance. Having some form of filter for unaffected cells will be crucial to filter out noise from the dataset. Additionally, a better translation needs to be made. Usage of sample level embeddings can be used, but you might lose the single-cell level data. The MTL shows potential, but it needs a decent model to improve, which is what is mainly missing in this research.

This research shows some potential in the classification of

AD severity measures using scRNA-seq data. If these models are improved, we can find patterns in certain genes that are responsible for the development of AD throughout the brain. This could help in understanding the cellular based mechanisms responsible for the development of AD, which could lead to finding a potential cure.

5 Acknowledgments

The results published here are in whole or in part based on data obtained from the AD Knowledge Portal. Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), R01AG17917 (MAP), R01AG30146, R01AG36042 (5hC methylation, ATACseq), RC2AG036547 (H3K9Ac), R01AG36836 (RNAseq), R01AG48015 (monocyte RNAseq) RF1AG57473 (single nucleus RNAseq), U01AG32984 (genomic and whole exome sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG46161(TMT proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute (genomic). Additional phenotypic data can be requested at www.radc.rush.edu. Research reported in this work was partially or completely facilitated by computational resources and support of the Delft AI Cluster (DAIC) at TU Delft (RRID: SCR_025091), but remains the sole responsibility of the authors, not the DAIC team.

6 Responsible Research

6.1 Usage of data

This research used the data of 450 individuals from the ROSMAP dataset. Given that all of the AD severity measures, as well as the single-cell gene expression data are only available post mortem, all of these individuals have passed away and donated their brains to science. This means data like this is very sensitive, and due to that fact, we needed to request access to the dataset with a detailed description of the purpose. Additionally, we had to accept terms that this data can only be used for the research purpose and that we are not allowed to share it. Finally, all data must be destroyed after the project is finished. We have kept to these terms. The data has only been on our personal machines, or on the Delft AI cluster. All copies on our personal machine have been deleted.

Furthermore, the dataset also contained additional metadata about the donors. 3 of these we used in our research, those being the age, sex and APOE genotype. The dataset also contained other information like race and years of education. An argument can be made since these are all connected to AD, instead of using them as prediction targets, you can use them as additional input. Very early iterations of the models did use this, but when analyzing these models, we found that the model almost exclusively used this metadata to classify AD, while almost not using the gene expression data. This can obviously lead to biasses where the model would for example really factor in the years of education without any additional context leading to very biased models. Additionally, since a big part of this research is about using the gene expression data, the choice was made to not include any metadata as additional input. As for the selection of the additional task, APOE genotype and age are clear, as they have obvious connection to AD. As for the argument for the usage of sex, there is very strong evidence that women are more likely to develop AD [17], additionally, the gene expression difference between male and female is quite different due to their difference in X and Y chromosome and different hormonal influences. This means it could be an interesting target for gene expression data to predict while also being connected to AD.

6.2 Machine learning in healthcare

The main usage of this model was to find out if machine learning can find some pattern in gene expression to quantify AD severity measures. If these patterns can be found, we can get more insight into the cellular level processes that cause AD and for example find genes that increase or decrease the risk of AD. From this, we can find out if certain individuals are more or less likely to develop AD. A potential issue with this is that healthcare insurers can also find out about this, and if they base their costs on patient data, they can use this information to charge certain people way more as they have a higher risk of developing AD, meaning they will require more care. The purpose of this research is not that. This data should only be used to help develop accessible treatment options.

Furthermore, if these models can perform at a high enough level, they can replace the tasks of diagnosis currently in the hands of neuropathologist and neuropsychologist that currently determine these AD severity measures. Models like this are perfect for finding patterns in huge dataset at speeds far exceeding human skills. Models like this however, are not humans, meaning they cannot be held responsible for mistakes they made. This is why ML should only be used as a tool for doctors to feed in a lot of data, and let the models pick out certain data that the doctor needs to investigate more closely. Models like this should never be the sole diagnosis giver and additional confirmation is always needed. Even a model with 99% accuracy, still makes mistakes, and its the job of the doctor to find and fix these mistakes.

References

[1] Alzheimer's Association. 2023 alzheimer's disease facts and figures. *Alzheimer's Dementia*, 19(4):1598–1695, 2023.

[2] 2024 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 20(5):3708–3821, May 2024. Epub 2024 Apr 30.

[3] National Institute of Aging. What causes alzheimer's disease? *National Institute of Aging*, 2025.

[4] Bong-Hyun Kim, Kijin Yu, and Peter C W Lee. Cancer classification of single-cell gene expression data by neural network. *Bioinformatics*, 36(5):1360–1366, 10 2019.

[5] Kazi Ferdous Mahin, Md. Robiuddin, Mujahidul Islam, Shayed Ashraf, Farjana Yeasmin, and Swakkhar Shatabda. Panclassif: Improving pan cancer classification of single cell rna-seq gene expression data using machine learning. *Genomics*, 114(2):110264, 2022.

[6] Dezhi Huang, Naya Ma, Xinlei Li, Yang Gou, Yishuo Duan, Bangdong Liu, Jing Xia, Xianlan Zhao, Xiaoqi Wang, Qiong Li, Jun Rao, and Xi Zhang. Advances in single-cell RNA sequencing and its applications in cancer research. *J Hematol Oncol*, 16(1):98, August 2023.

[7] Gilad Sahar Green, Masashi Fujita, Hyun-Sik Yang, Mariko Taga, Anael Cain, Cristin McCabe, Natacha Comandante-Lou, Charles C. White, Anna K. Schmidtner, Lu Zeng, Alina Sigalov, Yangling Wang, Aviv Regev, Hans-Ulrich Klein, Vilas Menon, David A. Bennett, Naomi Habib, and Philip L. De Jager. Cellular communities reveal trajectories of brain ageing and Alzheimer's disease. *Nature*, 633(8030):634–645, September 2024.

[8] Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, July 1997.

[9] Piu Upadhyay and Sumanta Ray. A regularized multitask learning approach for cell type detection in single-cell rna sequencing data. *Frontiers in Genetics*, Volume 13 - 2022, 2022.

[10] Xin Zhang, Le Gao, Zhimin Wang, Yong Yu, Yudong Zhang, and Jin Hong. Improved neural network with multi-task learning for alzheimer's disease classification. *Heliyon*, 10(4):e26405, 2024.

[11] T. Verlaan, G. Bouland, A. Mahfouz, and M.J.T. Reinders. scagg: Sample-level embedding and classification of alzheimer's disease from single-nucleus data. *bioRxiv*, 2025.

[12] Alejandra P. Perez-Gonzalez, Aidee Lashmi Garcia-Kroepfly, Keila Adonai Perez-Fuentes, Roberto Isaac Garcia-Reyes, Fryda Fernanda Solis-Roldan, Jennifer Alejandra Alba-Gonzalez, Enrique Hernandez-Lemus, and Guillermo de Anda-Jauregui. The rosmap project: aging and neurodegenerative diseases through omic sciences. *Frontiers in Neuroinformatics*, Volume 18 - 2024, 2024.

[13] Heiko Braak, Kelly Del Tredici, Udo Rüb, Rob A.I de Vos, Ernst N.H Jansen Steur, and Eva Braak. Staging of brain pathology related to sporadic parkinson's disease. *Neurobiology of Aging*, 24(2):197–211, 2003.

[14] H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, September 1991.

[15] *ROSMAP clinical codebook*.

[16] S. S. Mirra, A. Heyman, D. McKeel, S. M. Sumi, B. J. Crain, L. M. Brownlee, F. S. Vogel, J. P. Hughes, G. van Belle, and L. Berg. The consortium to establish a registry for alzheimer's disease (cerad). part ii. standardization of the neuropathologic assessment of alzheimer's disease. *Neurology*, 41(4):479–486, Apr 1991.

[17] Stefania Ippati, Lars Matthias Ittner, and Yazi Diana Ke. Chapter 1 - sex differences in alzheimer's disease animal models. In Maria Teresa Ferretti, Annemarie Schumacher Dimech, and Antonella Santuccione Chadha, editors, *Sex and Gender Differences in Alzheimer's Disease*, pages 3–22. Academic Press, 2021.

[18] National Institute of Aging. Alzheimer's disease fact sheet. *National Institute of Aging*, 2025.

[19] Lindsay A. Farrer, L. Adrienne Cupples, Jonathan L. Haines, Bradley Hyman, Walter A. Kukull, Richard Mayeux, Richard H. Myers, Margaret A. Pericak-Vance, Neil Risch, and Cornelia M. van Duijn. Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease: A meta-analysis. *JAMA*, 278(16):1349–1356, 10 1997.

[20] Qi Wang, Kewei Chen, Yi Su, Eric M. Reiman, Joel T. Dudley, and Benjamin Readhead. Deep learning-based brain transcriptomic signatures associated with the neuropathological and clinical severity of alzheimer's disease. *Brain Communications*, 4(1):fcab293, 12 2021.

[21] Minghui Wang, Won-min Song, Chen Ming, Qian Wang, Xianxiao Zhou, Peng Xu, Azra Krek, Yonejung Yoon, Lap Ho, Miranda E. Orr, Guo-Cheng Yuan, and Bin Zhang. Guidelines for bioinformatics of single-cell sequencing data analysis in Alzheimer's disease: review, recommendation, implementation and application. *Molecular Neurodegeneration*, 17(1):17, March 2022.

[22] Fabien Dachet, James B. Brown, Tibor Valyi-Nagy, Kunwar D. Narayan, Anna Serafini, Nathan Boley, Thomas R. Gingeras, Susan E. Celniker, Gayatry Mohapatra, and Jeffrey A. Loeb. Selective time-dependent changes in activity and cell-specific gene expression in human postmortem brain. *Scientific Reports*, 11(1):6078, March 2021.

[23] Delft AI Cluster (DAIC). The delft ai cluster (daic), rrid:scr_025091, 2024.

[24] Bernardino Romera Paredes, Andreas Argyriou, Nadia Berthouze, and Massimiliano Pontil. Exploiting unrelated tasks in multi-task learning. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 951–959, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[25] Tra-My Vu, Vincent Hervé, Anosha Kiran Ulfat, Daniel Lamontagne-Kam, and Jonathan Brouillette. Impact of non-neuronal cells in Alzheimer's disease from a single-nucleus profiling perspective. *Frontiers in Cellular Neuroscience*, 17:1208122, 2023. eCollection 2023. Review Article.

[26] Tamim Abdelaal, Soufiane Mourragui, Ahmed Mahfouz, and Marcel J T Reinders. Spage: Spatial gene enhancement using scrna-seq. *Nucleic Acids Research*, 48(18):e107–e107, 09 2020.

A Usage of LLM's

LLM's were mainly used in the coding and data analysis process to translate our idea's to python code. Some examples of prompts used:

Ordinal classification "For a model in python I am using ordinal classification, but I believe my lossfunction is not working correctly, what loss function should I be using and what is the python equivilent?"

"I am not fully sure if my ordinal classification pytorch model is correct. Could you write me write me a model that has 50 input features, 2 hidden layers and 4 output features?"

"My model is definitely overfitting:

this is the data at the first epoch: Epoch 1/15 Avg TRAINING loss for task 0: 0.496269

Avg TRAINING accuracy for task 0: 38.7Task 0 [coral]: Loss=0.5806, Acc=29.2Avg loss: 0.580583

And this at the final (15th) epoch: Epoch 15/15 Avg TRAINING loss for task 0: 0.081608

Avg TRAINING accuracy for task 0: 92.3Task 0 [coral]: Loss=1.0197, Acc=30.9Avg loss: 1.019730

What could be the causes of this? I have some idea's of my own, but would like to hear your opinion"

Translation part "in python, I have a list of a bunch of items, and I want to transform it into a dict with all the items and their count of occurances, can you make that for me?" "and how can I get the item with the highest count of of a dict?"

Data processing "I want to make an UpSet plot out of this data. Can you convert this to a csv with the green cells being true and white being false, and with the acc and std being split?"

"Can you also write a python code section that plots the braak acc with the standard deviation for the categories?"

"I ran multiple Multi Task Learning models and then compared accuracies on the models with various tasks, now I want to make a plot out of this, what would be a good way of showcasing this data?"

"Okay the accuracy is just on 1 'main' task, with the other tasks being there to potentially help the other one. Ive seen something like an UpSet plot, but that seems a lot more catered to showing true counts, and not really accuracies, what would be something similar to that, esspecially for the x-axis"

"I have the following file with the following dataformat:

idsBraak: 'R5693901': 2: 200, 1: 170, 0: 67, 3: 5, [MORE DATA HERE]

how can I read this data and get out all the dicts?"

"Can you write a function that parses that data from a file/"

"I have 5 of these dicts with slightly different values and I want to merge them into one where all the values are added. Can you write some pythong code that does this?

[RAW DATA]"