Chapter 4. Results

A step by step process was followed for building the models with best estimators of the diagnosis for each of the models described in Chapter 3. To make this process replicable, the code was included in a toolset called `open-dementia-reports`, and the code was versioned. The necessary artifacts for using the pipeline are:

- dataset files: `adni_df.csv` was the merged file from the singular files provided by ADNI, and `nacc_df.csv` was the file supported by `NACC` with no changes.
- configuration file: `config.json` that includes datasets description and model definition: dataset built on, diagnostic values, and the predictors included. Attached at Appendix C an example.
- variables files: for each dataset the variables were categorized into `Categorical` and `Numerical`, and described: missingness percentage, short descriptor, source (ex. `ADNIMERGE.csv`), potentially clinically relevant, exclusion criteria if so, and descriptive statistics.

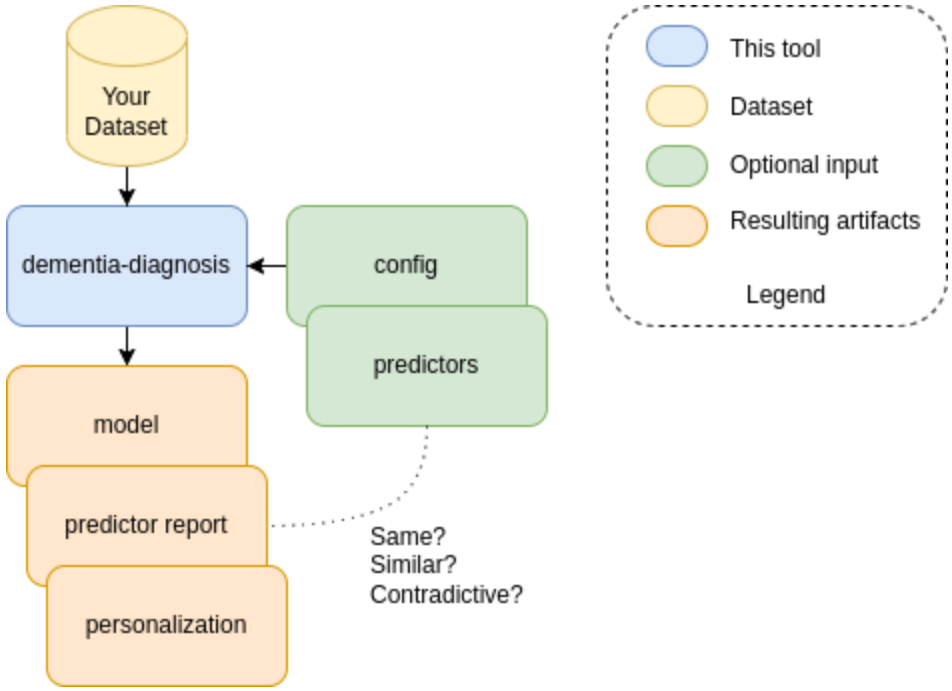

Figure 1: Processing tool.

The `config.json` and the `variables.csv` files are added in the supplementary material. The code for reproducing this work can be found online.[1] The code is written in the `python` programming language because of its extensive library support. The two main functionalities are shown in the reports `Report-Build-Model.ipynb` and `Report-Explain.ipynb`. The first one does the testing for finding the best pipeline, while the second one can be used for explaining the impact of the predictors.

[1] https://github.com/DorenCalliku/open-dementia-reports

4.1. Datasets

Participants The visits for each subject were extracted from the dataset files based on their diagnosis. Participant characteristics for each model are described below. Some observations:

- Both datasets contain the HC-MCI-AD spectrum,
- NACC dataset contains more patients with more visits,
- The gender distribution changes depending on the disease, with DwMD/DLB as the most skewed,
- The age range is similar, except for the patients with FTLD in the NACC dataset,
- The education levels are similar for the datasets.

Table 1: Dataset descriptions.[2]

Diagnosis	Subj. (Visits)	Gender (F %)	Age (std)	Education (std)	Models
ADNI					
AD	2426 (821)	1037 (42.75%)	74.29 (7.36)	15.49 (2.88)	M1, M2
HC	3968 (984)	2096 (52.82%)	72.96 (6.27)	16.52 (2.58)	M1, M2
MCI	4994 (1241)	1998 (40.01%)	72.9 (7.45)	15.99 (2.82)	M1, M2
NACC-base					
AD	56907 (20930)	30711 (53.97%)	76.95 (9.8)	15.4 (7.11)	M3, M4
HC	78217 (19743)	51075 (65.3%)	73.82 (10.19)	16.29 (6.09)	M3, M4
MCI	4284 (2894)	2246 (52.43%)	72.54 (10.77)	15.3 (5.58)	M3, M4
M3					
DLB	4698 (2082)	1073 (22.84%)	73.8 (8.48)	16.08 (6.25)	M3
VaD	3708 (2016)	1928 (52.0%)	79.04 (8.77)	15.69 (8.56)	M3
FTD	6174 (2521)	2582 (41.82%)	66.03 (9.45)	17.01 (11.17)	M3
M4					
DwMD	5852 (2703)	1668 (28.5%)	72.84 (8.6)	16.26 (7.54)	M4
VaD	3708 (2016)	1928 (52.0%)	79.04 (8.77)	15.69 (8.56)	M4
FTLD	6251 (2552)	2614 (41.82%)	66.01 (9.43)	17.01 (11.21)	M4

[2]Both models M1 and M2 of ADNI represent the same data, but for NACC there is a difference in the data that models M3 and M4 use: M3 uses the data for DLB and FTD, and M4 uses DwMD and FTLD, following the grouping after Chapter 2.

Exploratory Data Analysis Exploratory data analysis (EDA) allows for an observation on how the variables are distributed, and it can provide some insight on what to expect. As we have suggested in Chapter 2, there are several variables that are expected to differentiate between the disorders, and exploring how they are represented in the dataset can provide a better insight on how well they might do, and how much is the actual method of explanation adding to the EDA.

ADNI As mentioned in Chapter 2, there are several factors that can differentiate the diseases in the HC-MCI-AD spectrum. The development of AD since HC requires a drop in the cognitive functioning, and a neuro-degeneration expressed initially at the middle brain, and then spreaded in the other layers. There exists no one test that can differentiate the diseases properly, even though the CDR (Clinical Dementia Rating scale) responds consistently well. In the ADNI dataset there are several general purpose tests (e.g. CDRSB, ADAS13, MMSE, MOCA) and more specific tests (e.g. AVDELTOT, EcogSPMEM tests, ADNI_MEM memory test). Additionally there are also measurements of the brain regions in terms of volume. How these tests and brain regions are separated for the HC-MCI-AD spectrum in the dataset can be seen below. The main observation is

NACC The NACC dataset contains a larger variation of the disorders (HC, MCI, AD, FTLD, DwMD, VaD). Having this high variation makes the graphical observations more difficult, but might provide direct inference on how one class is separable from the others. As it can be observed below: NACCAGE is a good separator for FTLD diseases, MOCATOTS helps for a good separation for HC, CFRAFT examinations (DRE/URS) help at separating FTLD-AD from the other diseases. The NACC dataset contains most of the data encoded in a categorical format. Below the description through a radar-plot of the scores of the predictors.

- Impairment: The group `{FTLD, DwMD, AD}` tends to be more impaired, mainly losing independence. `FTLD` tends to be more impaired in Language and behavior than the other diseases.
- Medical history: Presence of depression in the last two years comes hand in hand with most of the diseases. Incontinence adds to the problems of `DwMD`. Interesting is the presence of Diabetes for patients with `VaD`.
- Parkinsonism and tremor-related symptoms: These symptoms are highly present in `DwMD`. Additionally, `GAIT` seems to be impaired also for `VaD`.
- Cardiovascular: `CBSTROKE` is as expected a risk for `VaD`. Still, it can be observed that the patients of other diseases do have similar problems with `HYPERTEN` and `HYPERCHO`. `CVDCOG` seems to target specifically `VaD` subgroup.
- Neuroimaging: Presence of `HIPPATR` (hippocampus atribution) for diseases like `FTLD` and `AD`. Additionally, the presence of `CVDIMAG` suggests the presence of `VaD`.

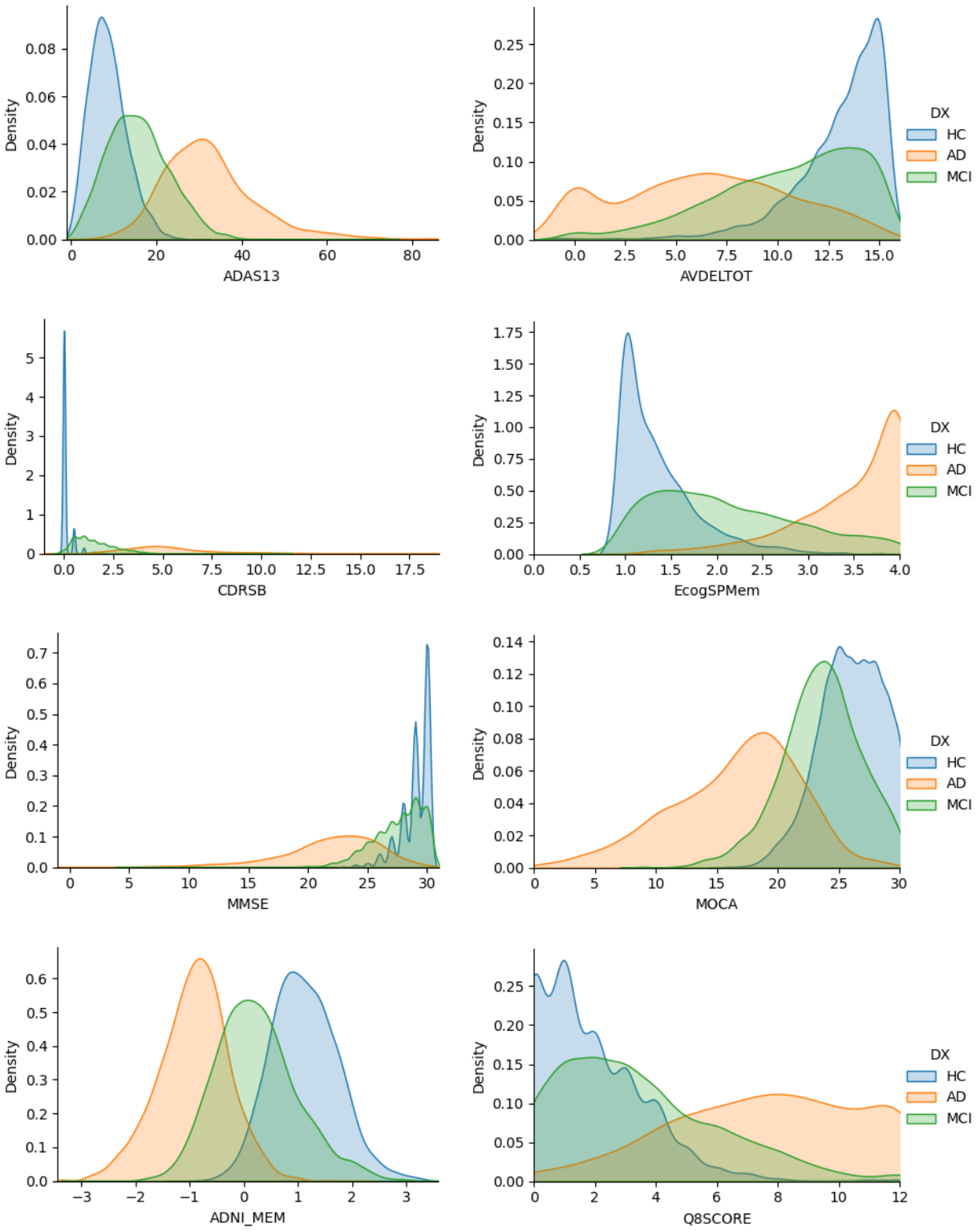

Figure 2: Cognitive scores related to ADNI.

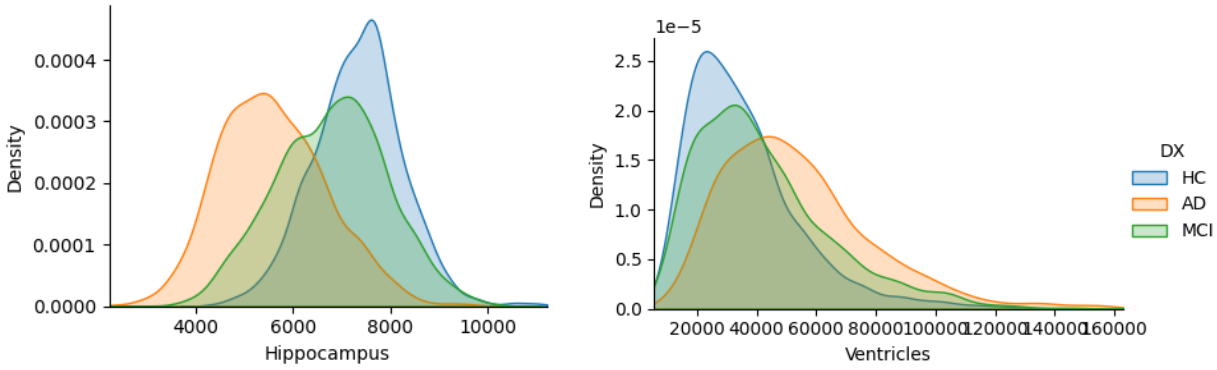

Figure 3: Hippocampus and Ventricles volumes for ADNI.

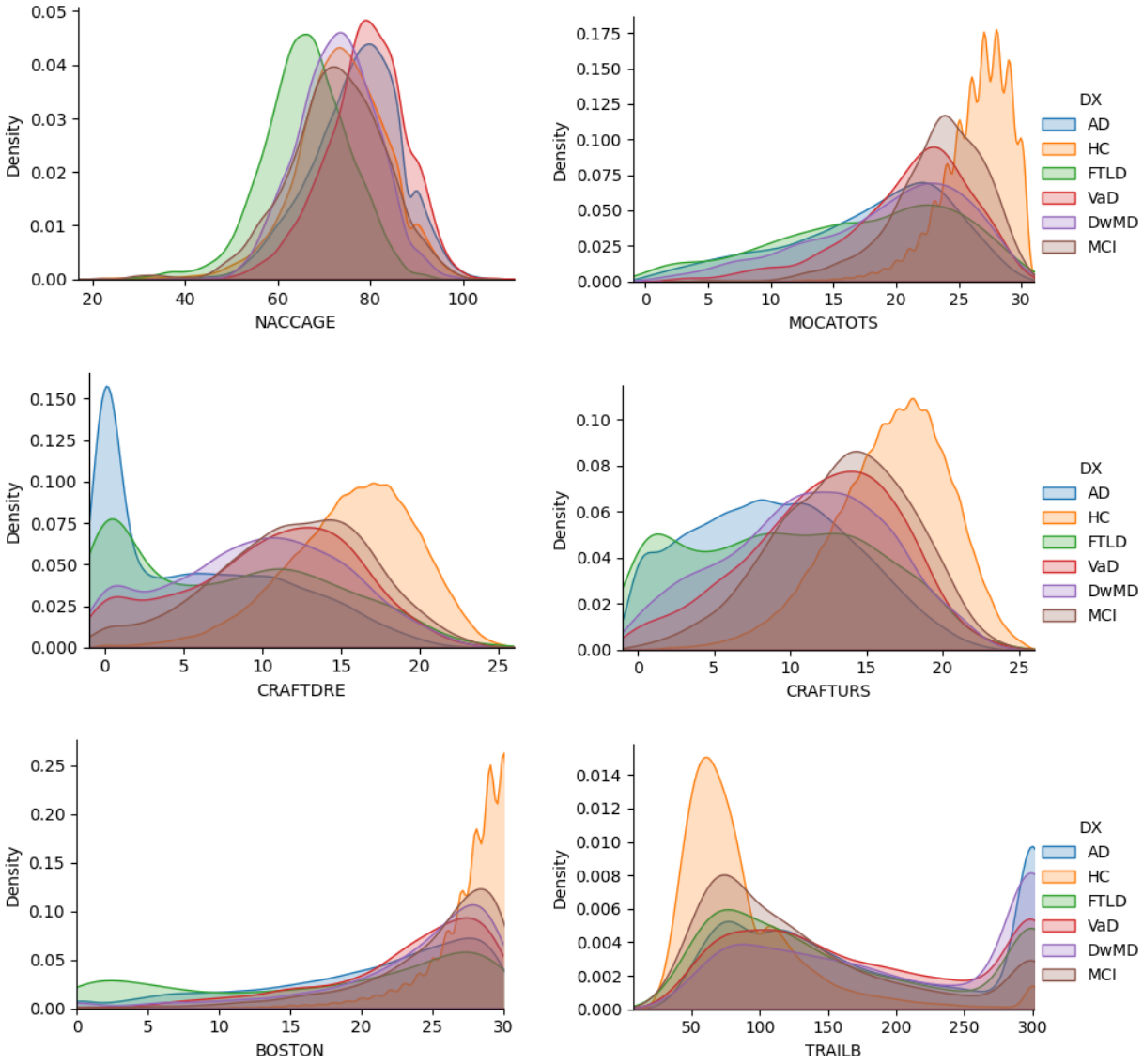

Figure 4: NACC numerical data samples.

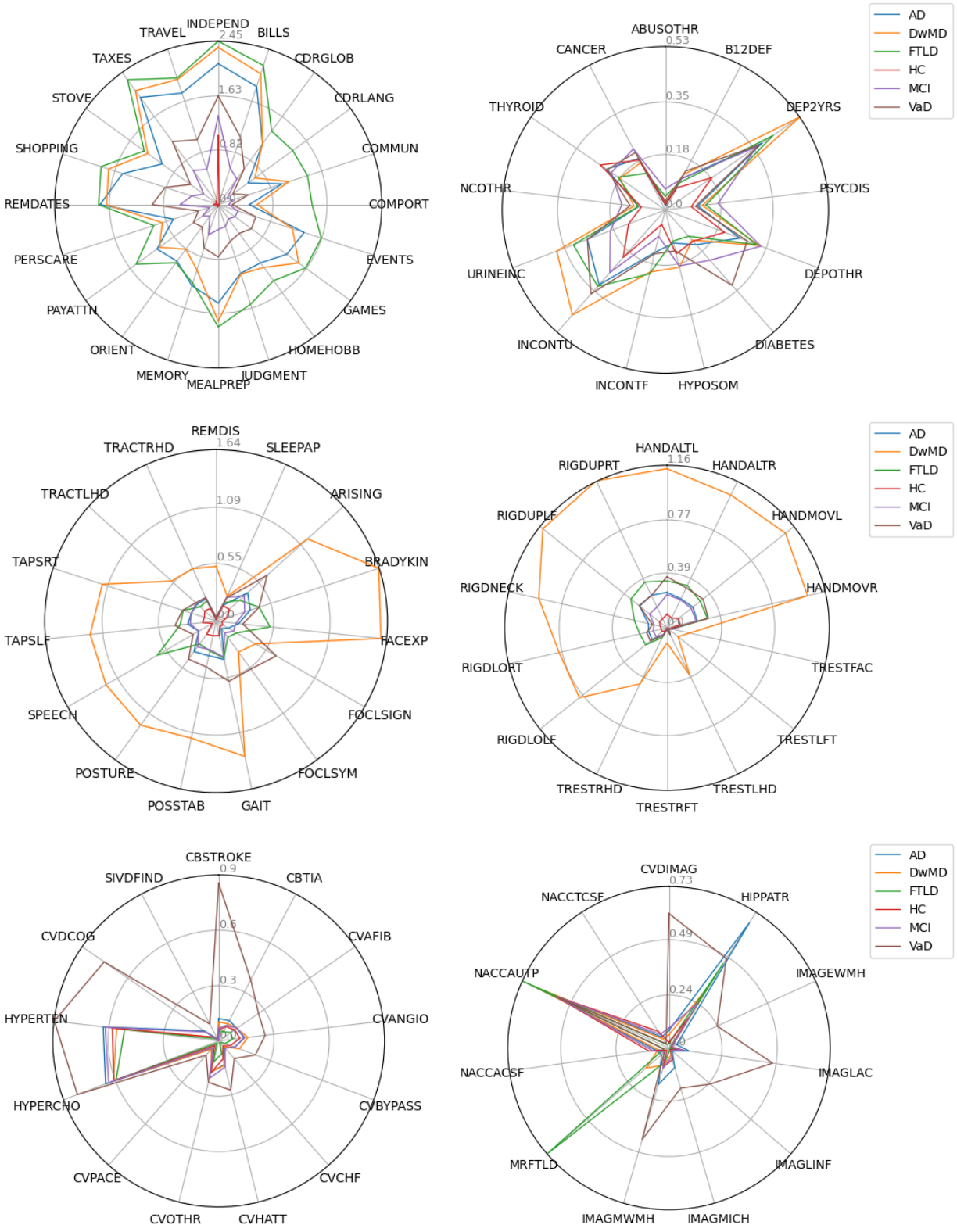

Figure 5: Impairment, medical history, parkinsonism symptoms, tremor-related symptoms, cardio-vascular history, and neuroimaging categorical data in the NACC dataset in that order.

Table 2: Relevant variables mentioned above, and their descriptors.

Variable	Category	Short descriptor	mean	std
ADNI				
CDRSB	impair	Clinical Dementia Rating - Sum of Boxes	1.93	2.97
Hippocampus	nimg	UCSF Hippocampus	6891.72	1221.92
Ventricles	nimg	UCSF Ventricles	38999.1	21183.7
ADAS13	npa	AD Assessment Scale 13	16.59	11.18
ADNI_MEM	npa	Memory summary score	0.45	1.01
AVDELTOT	npa	Recognition Score	11.03	3.81
EcogSPMem	npa	Study Partner ECog - Mem	2.1	1
MMSE	npa	Mini-Mental State Examination	27.11	3.77
MOCA	npa	Montreal Cognitive Assessment	23.33	4.74
Q8SCORE	npa	Score Component	3.61	3.21
NACC				
INDEPEND	demographics	Level of independence	1.54	0.89
NACCAGE	demographics	Subject's age at visit	74.61	10.25
CBSTROKE	health	Stroke	0.1	0.43
DEP2YRS	health	Active depression in the last two years	0.3	0.46
INCONTU	health	Incontinence — urinary	0.27	0.52
CDRLANG	impair	Language	0.29	0.65
COMPORT	impair	Behavior, comportment, and personality	0.26	0.64
CVDIMAG	nimg	Imaging evidence	0.08	0.27
IMAGLAC	nimg	Lacunar infarct(s)	0.09	0.28
IMAGMWMH	nimg	Moderate white-matter hyperintensity (CHS score 5-6)	0.13	0.34
BOSTON	npa	Boston Naming Test (30) — Total score	24.31	6.53
CRAFTDRE	npa	Craft Story 21 Recall (Delayed) Total story units recalled, paraphrase scoring	12.22	6.45
CRAFTURS	npa	Craft Story 21 Recall (Immediate) Total story units recalled, paraphrase scoring	13.78	5.63
MOCATOTS	npa	MoCA Total Raw Score — uncorrected	23.09	6.03
TRAILB	npa	Trail Making Test Part B — Total number of seconds to complete	124.22	82.01
BRADYKIN	physical	Body bradykinesia and hypokinesia	0.28	0.66

Variable	Category	Short descriptor	mean	std
GAIT	physical	Gait	0.3	0.67

4.2. Pipelines

Finding the better processing pipelines required testing the combination of all the methods for selecting the best preprocessing steps, best models, and tuning relevant parameters. The pipelines were put under a grid-search, as suggested in Chapter 3. All the results were cross-validated through `StratifiedGroupKFold`, with grouping based on subject ID (`RID` for the ADNI dataset, `NACCID` for the NACC dataset). The metric used was `f_beta score` with `beta=2` for giving more importance to sensitivity than precision. The results are stratified cross-validated with CV=4, for keeping a 75%-25% train-test ratio.

Some of the processing units do not allow missing values, as mentioned above: vanilla logistic regression, vanilla random forest, transform, unsupervised feature selection. While the boosted models like `LGBMClassifier` and `XGBClassifier` allow for missing values, the `RandomForestClassifier` does not allow. Additionally, all the ADNI dataset processing included some encoding before usage (`OrdinalEncoder` was used, as it did not expand the dataset as much as `OneHotEncoder`).

While having all the features is important, having only a subset of the features allows for keeping the model less sparse and more concentrated. For that the models were also tested with a supervised feature selection of 50 predictors, based on the `SelectFromModel(LGBMClassifier(n_estimators=50), max_features=50)`. This allows a comparison with the full features, and this interaction tends to show a decrease in the `f-beta-score` for both NACC and ADNI models.

Overview The results about the main pipelines are shown in the table below. From these results the following can be inferred:

- `vanilla-lgbm` pipeline does better than the other vanilla models overall.
- Adding classic preprocessing steps to the `vanilla-lgbm` does not significantly improve the scores, including sampling, imputing, sampling, and feature selection.
- There are no major differences between the minimal pipelines of different modeling. So, apparently there are no major differences between logistic regression, random-forest, and lgbm classifier once the scale is big enough. This can be for multiple reasons, but mainly it is the impact of on-to-all reflecting in the measurements.

Table 3: Performance f-beta-score of several of the pipelines.

Models	M1	M2	M3	M4
LGBM models				
vanilla-lgbm	**0.876444**	**0.886692**	**0.705914**	**0.701612**
unsupervised_fs	0.862309	0.876208	0.683106	0.675163
transform	0.87927	0.885749	0.701902	0.704858
sampling	0.876679	0.886103	0.703162	0.704076
imputation	0.87609	0.885043	0.701431	0.70083
Other models				
vanilla-rf	0.876915	0.881509	0.674846	0.671977
vanilla-lr	0.861366	0.855123	0.66643	0.670875

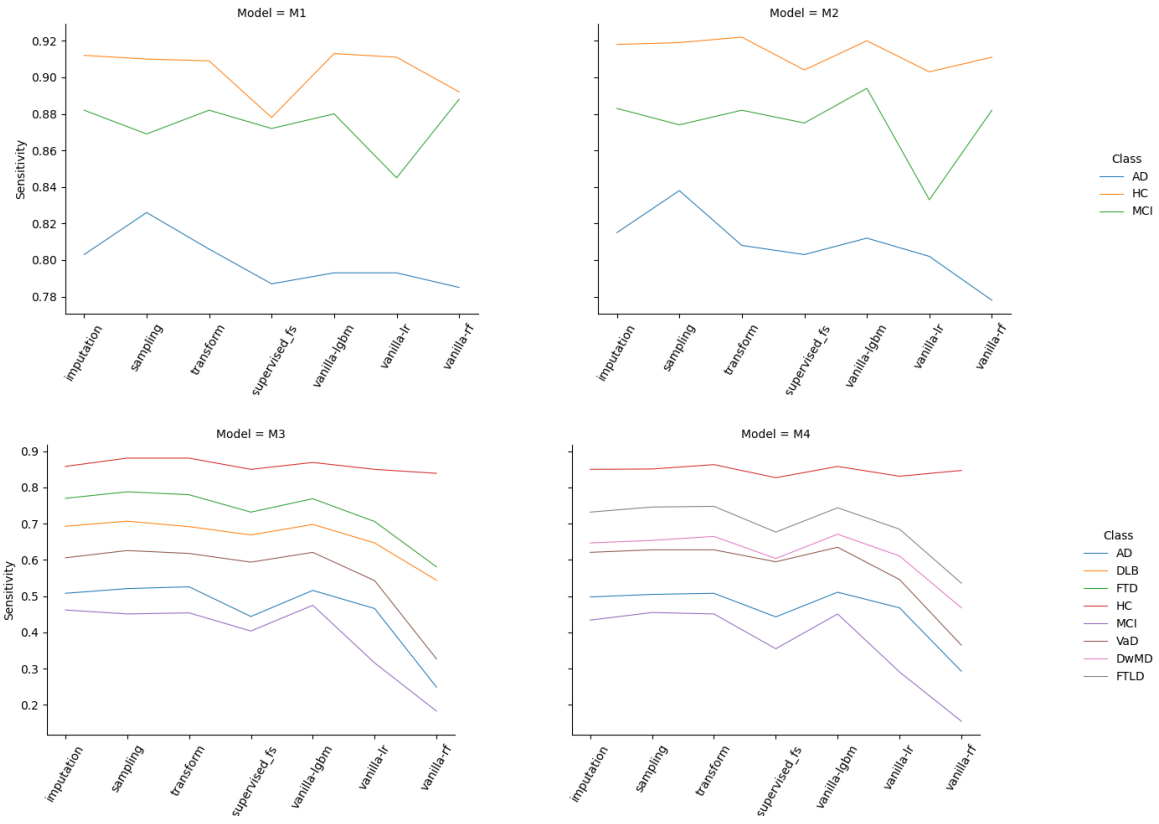

Figure 6: ADNI and NACC 'Sensitivity' scores for each class. As it can be observed, the better models are the LGBM models, as they remain sensitive for MCI in ADNI models and for MCI and DwMD for NACC.

Best non-redundant models For the analysis the `Vanilla-LGBM` model was picked because of the results, and the lack of requirement for the preprocessing steps. To understand how the model behaves for each class below are the scores.

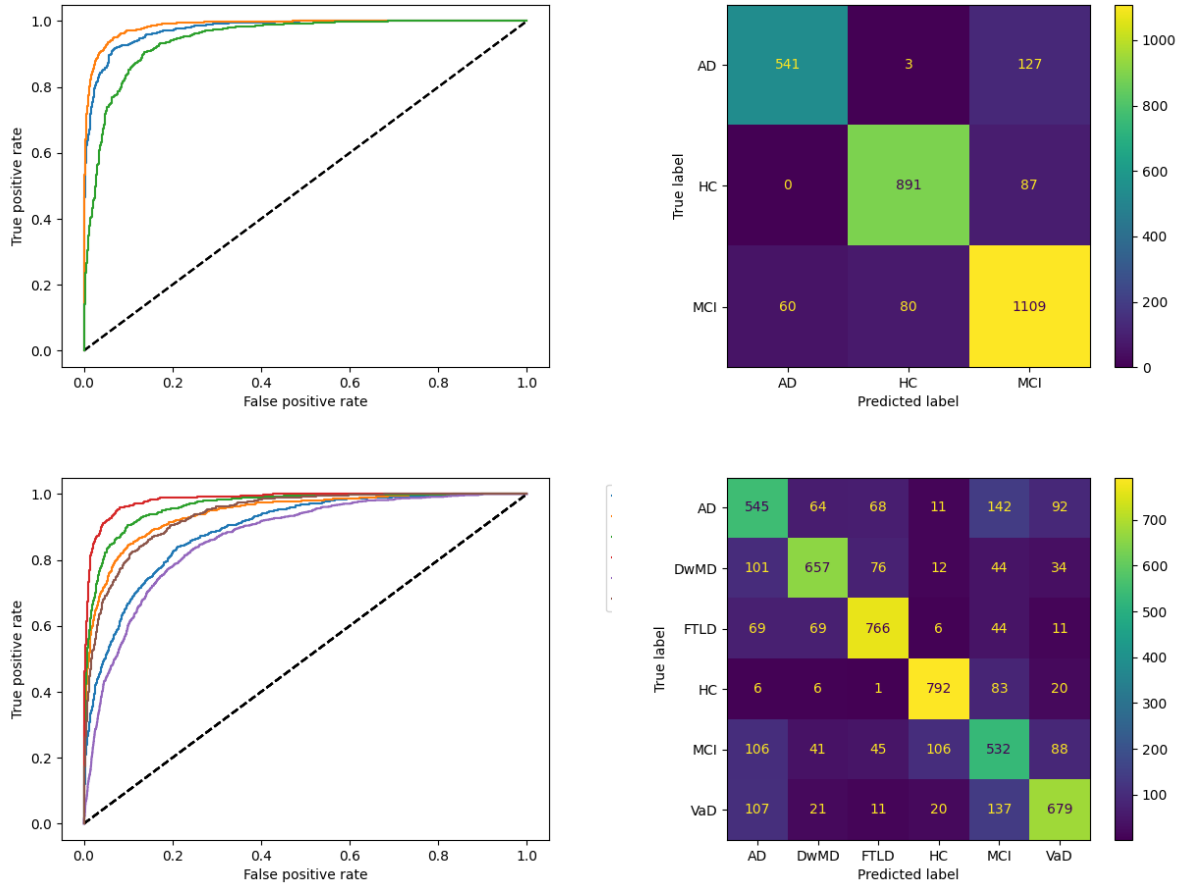

Figure 7: ROC curves and the confusion matrix for the M2-ADNI (first row) and M4-NACC (second row) vanilla-lgbm models. The ROC curves combine the true positives versus false positives, while the confusion matrix shows how the models do in terms of combined results. In the first row ADNI graphs, with green-MCI, blue-HC, orange-AD. In the second row, MCI-purple, AD-blue, VaD-brown, DwMD-orange, FTD-green, and HC-red.

Misclassifications These scores are close to the findings from the literature, and the models are learning those protocols, as shown below. Still, there is a good degree of `wrongly` classified cases. In the case of ADNI, the misclassification of AD-MCI is the most problematic, while for NACC the mis-classified cases are more heterogenous. To plot these misclassifications, the prediction probabilities of each case was checked, and it was observed whether there are some regularities in the distribution of probabilities. A subject with FTLD (`real`), for a prediction of being a healthy subject HC (`pred`), a comparison between the predicted value (`pred-val`) and the real value which was missed (`real-val`). Some observations:

- `MCI` tends to be the most mis-classified case for both ADNI and NACC.
- The three classes `AD-VaD-MCI` tend to overlap in misclassifications.
- If a healthy control has been misclassified, it tends to be either `MCI` or `VaD`.
- Mistakes of subjects with `FTLD` or `DwMD` predicted as `HC` are sparse, but strongly so.
- `DwMD` co-exists with `FTLD`, as differently from the other classes, tends to be mis-diagnosed as such.

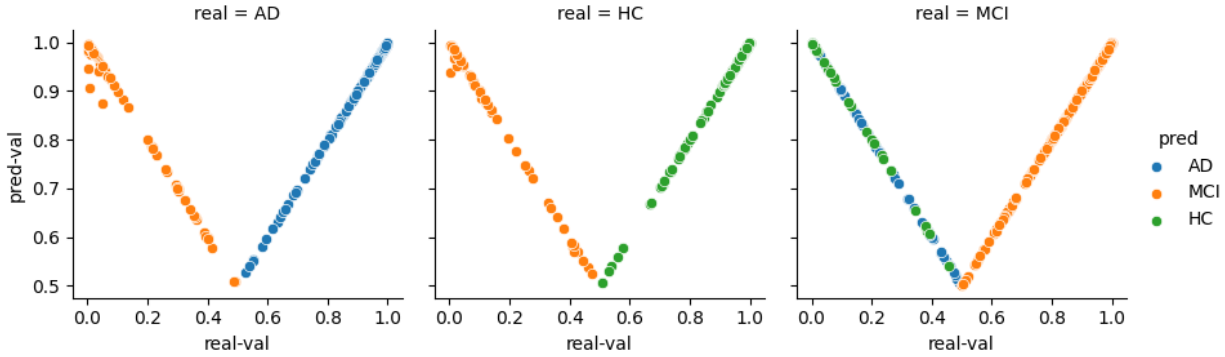

Figure 8: M2 distribution of mis-classifications.

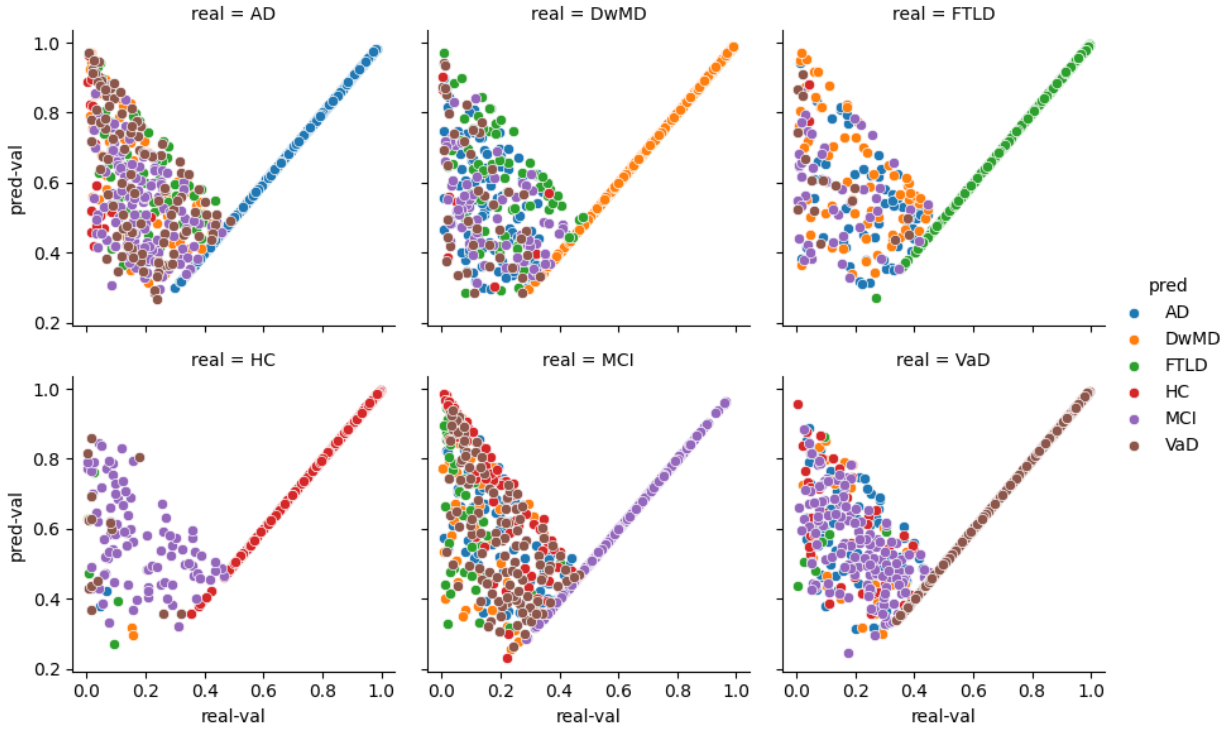

Figure 9: M4 distribution of mis-classifications.

4.3. Explainability

The global explainability for the two datasets was based on the `vanilla-lgbm` models. The global features can provide some insight on how the protocols are reflected in the scores, the interaction values show the combination of effects, and the local explainability shows direct implication in the cases.

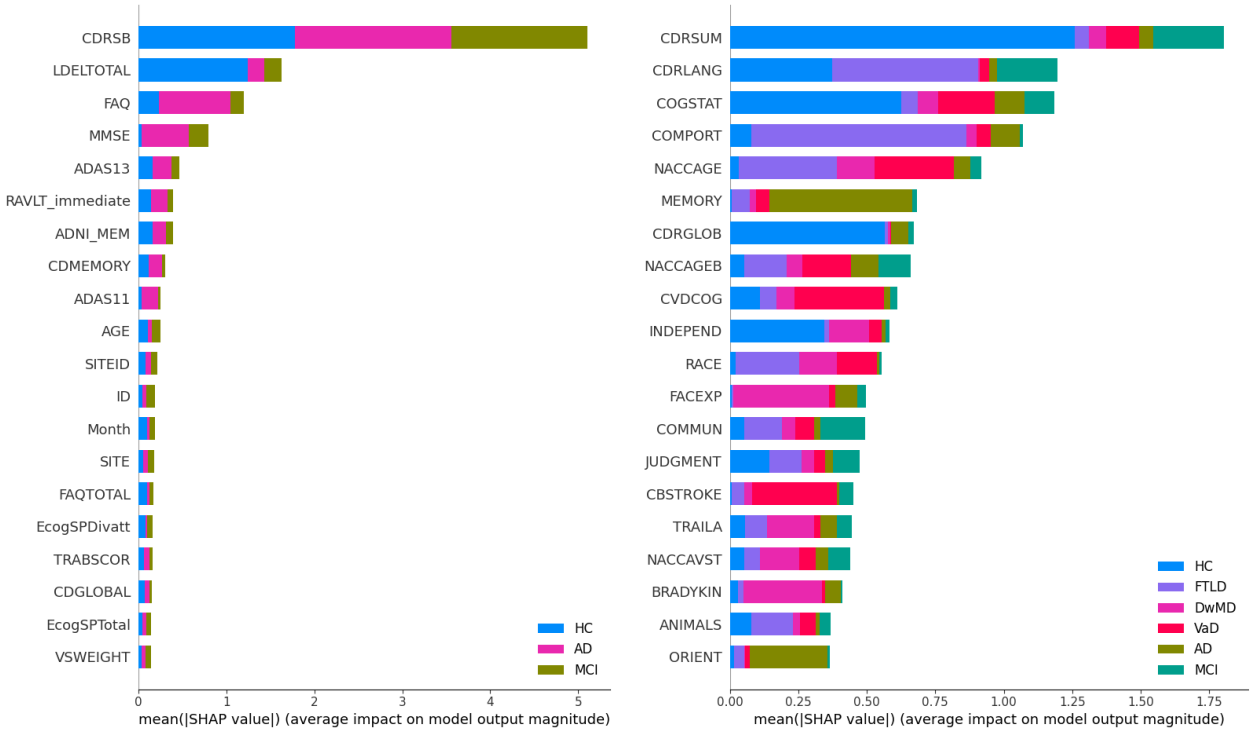

Figure 10: Feature importance for ADNI and NACC models respectively, ordered by impact.

In the plots above, the main biomarkers for the ADNI are `CDRSB`, `LDELTOTAL`, `FAQ`, `MMSE`, and `ADAS`. `LDELTOTAL` seems to be a good indicator of healthiness, and `FAQ` of AD. On the other hand, for NACC there is an integration of indicators. `CDRSUM` and `COGSTAT` seem important for indicating healthiness,`CDRLANG` (language), `COMPORT` (behavior), and `NACCAGE` are important for FTLD, cardiovascular history or risk expressed through `CVDCOG`, `CBSTROKE` and `HACHIN` are important for VaD, `MEMORY` is important for AD, `FACEXP`, `BRADYKIN`, and `INPEPEND` are important for DwMD, and MCI is distributed between the features. These elements reflect the criteria mentioned in Chapter 2. This suggests that the models have learnt as expected the interaction of the main features, and are able to differentiate between the diseases.

The importance can be seen distributed, but this does not provide a good understanding of how the scores of the tests impact the diagnosis. In the following this relation is observed for the classes tested in M2 and M4 as representative for ADNI and NACC.

4.3.1. ADNI

In the figure below, shapley values of ADNI through M2 are shown. As the HC-MCI-AD is a spectrum, it can be observed how each test behaves in the spectrum.[3] The following are notable impacts:

- `CDRSB`: high scores are predictors of AD, and low scores are predictors of HC.
- `LDELTOTAL`: high scores are strong predictors of HC.
- `FAQ`: high scores are predictors of MCI-AD.
- `MMSE`: high scores are predictors of HC-MCI, but the scores can overlap even in AD.
- `RAVLT_immediate`: high scores are predictors of HC.
- There are more factors involved in the recognition of a case of `MCI`, as it can be seen that the scores in the spectrum are more balanced than for the other two classes.
- Factors like `Site` and `Month` (of test) which should not impact the prediction come up as important for MCI prediction, suggesting some kind of not well-defined class, affectable from the artifacts.

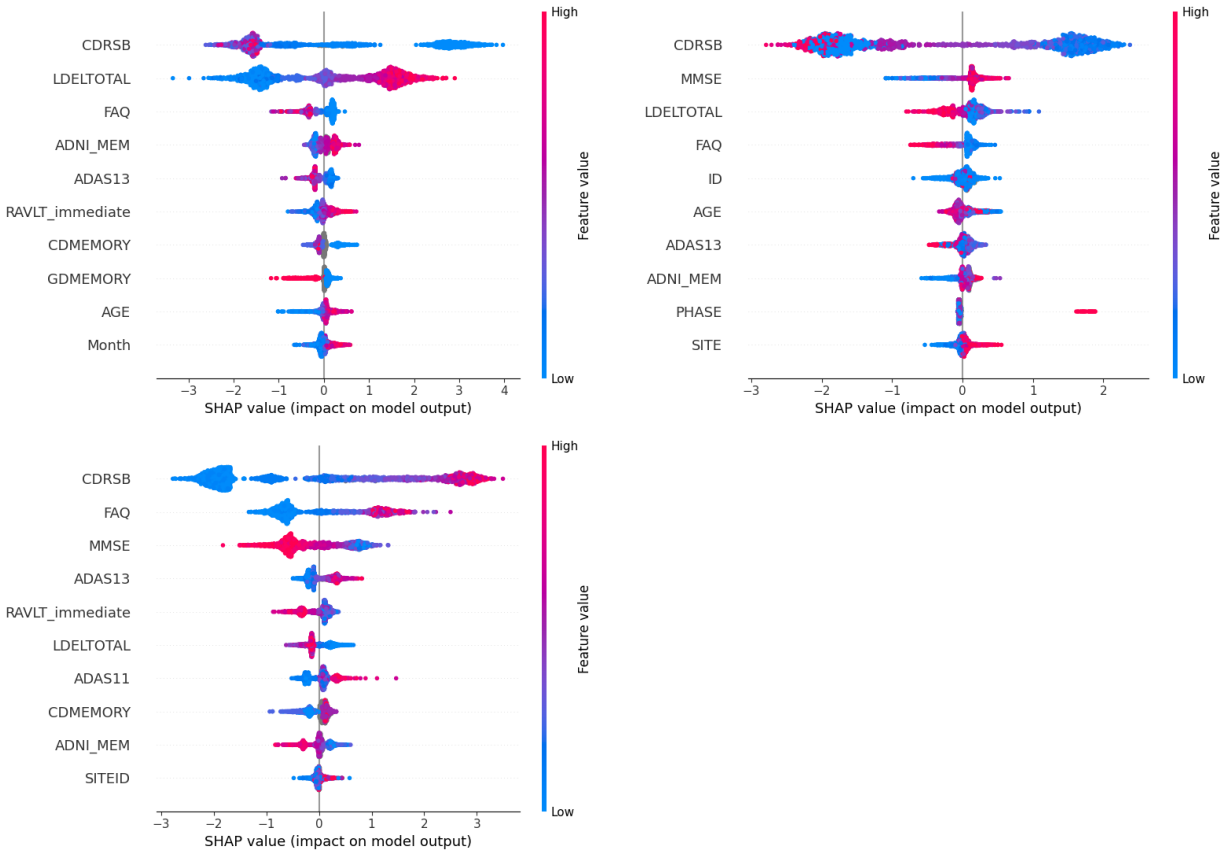

Figure 11: ADNI summary plot for HC, MCI, and AD respectively.

[3]How to read a summary plot: Color shows the real value where the blue suggests for lower values, while the shapley value is the level of importance of the factor in comparison to the others.

Single cases The implications of the predictions are checked through the waterfall plot.[4] The question that the plot is answering is: Does this subject show `MCI` patterns? And as it can be observed, for the HC and AD cases, the blue color suggests that these subjects are different from MCI. On the other hand, the plot on the right suggests that the scores `CDRSB = 2` and `FAQ = 0` strongly push for a prediction of MCI.

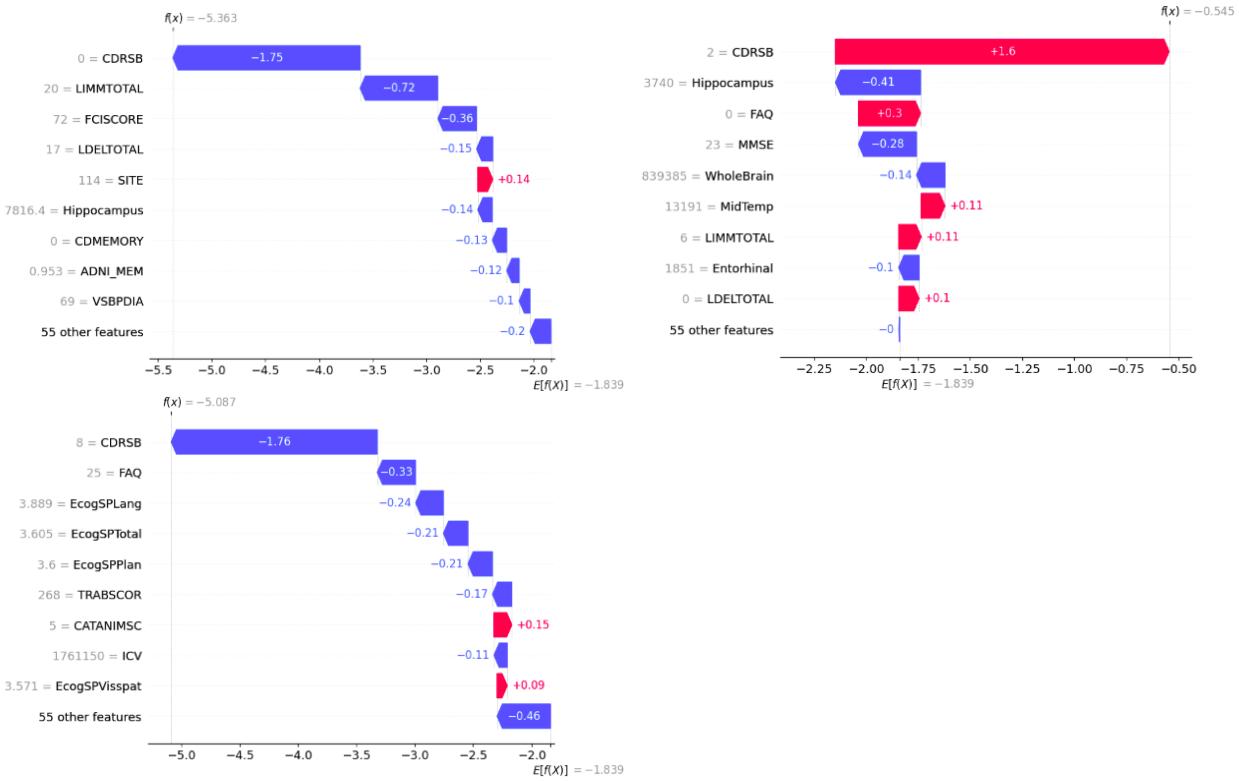

Figure 12: ADNI single case model decision descriptions for HC, MCI, and AD.

4.3.2. NACC

Differently from ADNI, the 6 classes of NACC make it more difficult for the model to settle on a small number of significant factors. As it can be seen below, all the classes (except `HC`) include more than 4 factors to be taken into consideration for making a decision.

- HC is the most well defined class with `CDRSUM` and `COGSTAT`.
- MCI summary is mostly related to negation of `AD-VaD` group and FTLD, as it is expressed clearly in the plot - low scores of `CDRLANG`, `CDRSUM`, `COMMUN`.
- AD is strongly related to `MEMORY` and `COGSTAT` scores, and excluding factors are FTLD factors like `COMPORT` (behavior), low `NACCAGE`, or DwMD factors like `SPEECH`.

[4] How to read a waterfall plot: The starting calculation measurement

- DwMD is dependent on the clinically defined symptoms like `FACEXP`, `RIGDUPRT`, `REMDIS`, `BRADYKIN`, and a neuropsychological test like `TRAILA`, and a clear preference for the `SEX` male.
- FTLD is defined mainly from `COMPORT` and `CDRLANG`, low `NACCAGE`, and some preference for `RACE`.
- VaD is defined by the late `NACCAGE`, presence of stroke (`CBSTROKE`, `CVDCOG`, `HACHIN`), and some less relevant predictors.

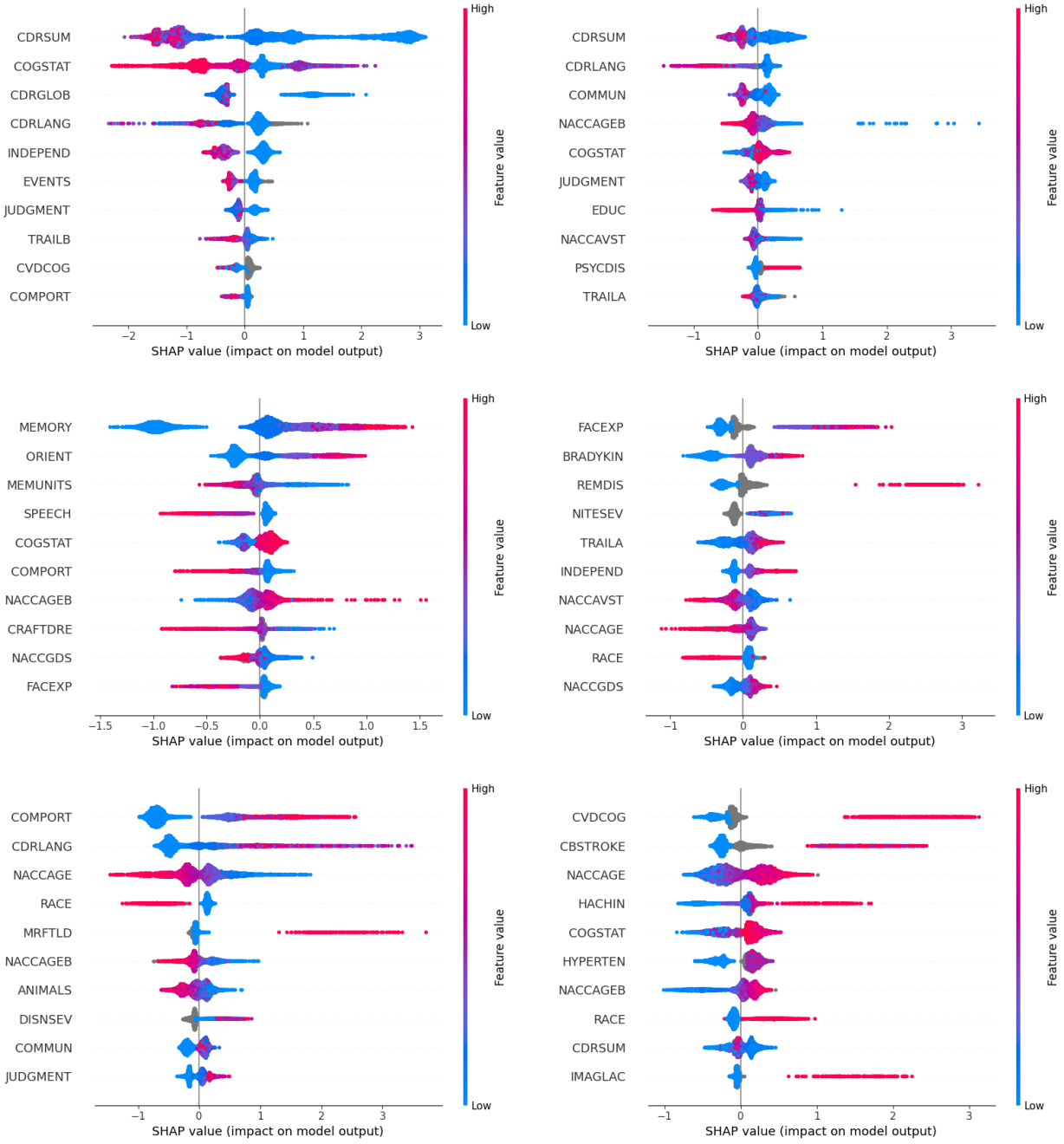

Figure 13: NACC summary plot for HC, MCI, AD, DwMD, FTLD, and VaD in that order.

Single cases The question that the plots below are answering: Do these subjects show AD patterns? The healthy subject, because of not having any problems with `MEMORY` and `ORIENT` is clearly not showing AD symptoms (check summary plot above of AD). In the other cases we can see how memory and orientation are generally impaired, making it difficult for the model to differentiate with the other disorders. The other subjects show AD patterns, even though not as clear as shown for the AD case. In these cases, a more specific analysis need to be run, for the data to be seen in reflection to the other `base_values`.

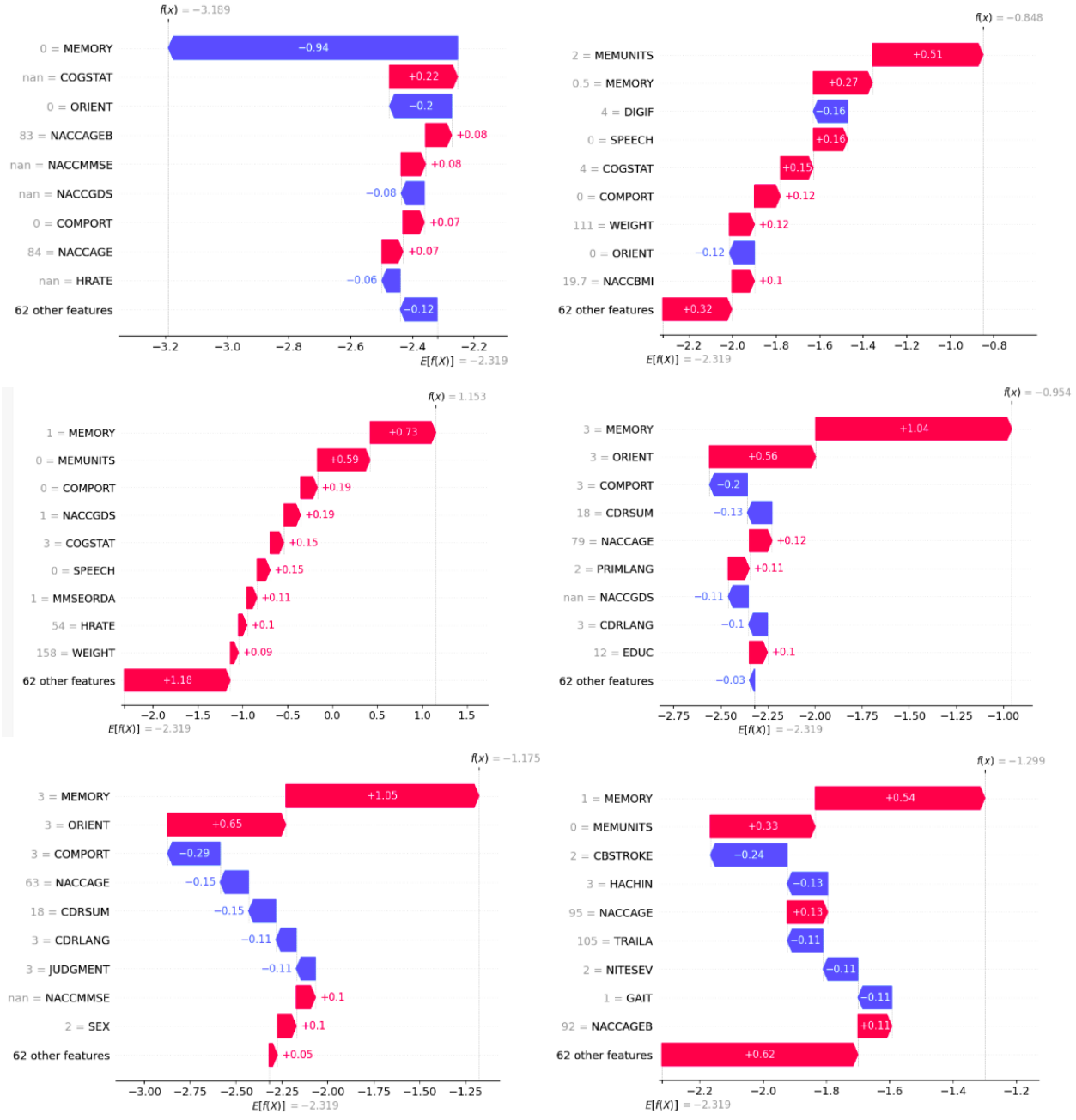

Figure 14: NACC single case model decision descriptions for HC, MCI, AD, DwMD, FTLD, and VaD.