Chapter 3. Materials and methods

3.1. Datasets

3.1.1. Alzheimer's Disease Neuroimaging Initiative

Markers for the diagnosis of Dementia have been based on open clinical datasets or local datasets. As mentioned by Kumar and colleagues, more than 60% of the research until 2019 has been built on the Alzheimer's Disease Neuroimaging Initiative (ADNI) and around 20% has been built using local datasets (Kumar et al. 2021). Most of the research has been built around ADNI because it contains a wide range of information, like clinical history, biomedical, neuroimaging, neuropsychiatric, and neuropsychological measurements. ADNI contains longitudinal multisite clinical data from patients of 55 research sites, funded by the National Institute of Health (NIH) and the industry. More information can be found at their official website (http://adni.loni.usc.edu/). The observational study has had three main cohorts, with an extension of the previous measurements in each cohort, and in this study only the cohorts ADNI-2 and ADNI-3 were included, so ADNI-1 and ADNI-GO were excluded. These two cohorts were selected because of the homogeneity of the data encoding.

The whole dataset is made of a total of 15171 visits of 2294 subjects. These subjects are diagnosed in one of the groups of Healthy Controls (no symptoms or memory complaints), Mild Cognitive Impairment (early or late onset), or Alzheimer's Disease patients. The dataset is heterogeneous as it includes demographics, genetics on the presence of APOE4 variation gene (genetics), neuroimaging information extracted through previous research on the variations of Magnetic Resonance Imaging techniques, neuropsychological assessment (npa), and more. For each of these features, there is a level of missingness that was handled later in the pipeline.

The level of information stored by the ADNI is extensive, and there was a need to select the relevant features before putting the data to a model. The files were downloaded from the ADNI website and categorized on their relevance. The baseline file was `ADNIMERGE` which has been observed extensively in the literature as the go-to selection of information (**???**). The other information was joined to the `ADNIMERGE` table through a left-join based on three columns: `RID` (patient ID), `VISCODE2`[1] (visit code based on months since first meeting), `ORIGPROT` (original protocol, so either ADNI-2 or ADNI-3). 80 files were initially included in the selection, and 54 files were excluded after exploration for one of the following reasons. For a complete list of the files and reasons of exclusion, check Appendix D, and for the script producing the merging of the data, check Appendix C.

[1] `VISCODE` in ADNIMERGE corresponds to `VISCODE2` in the other files.

Table 1: Files excluded from the analysis.

File type	Count	Exclusion reason
additional	16	Redundant, Detail, Minimal
co-participant	6	Minimal
diagnosis	6	Detail, Minimal
family-history	3	Redundant, Excluded
health-history	6	Detail, Excluded
impair	2	Minimal
medications	2	Redundant, Detail
nimg	2	Detail, Excluded
npa	7	Redundant, Detail, Excluded, Minimal
physical	3	Detail

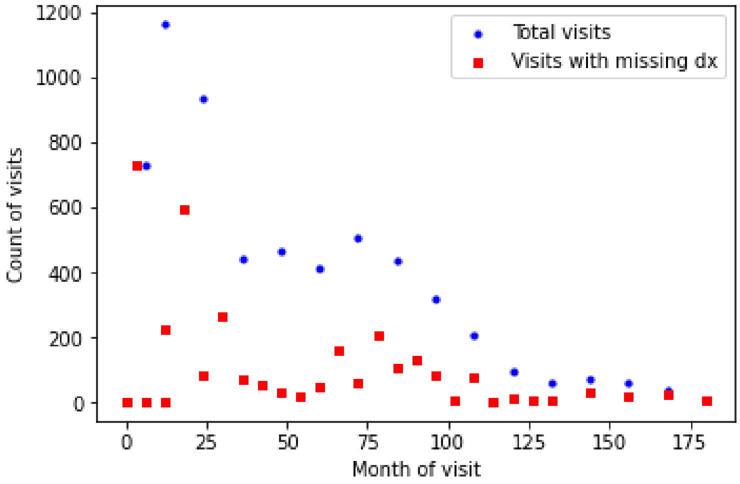

Figure 1: Missing diagnoses in ADNI.

- Detail: means that it is either a file that includes mostly specific information that might not have predictive value, or it is a file that contains mostly organizational information about the visit.
- Minimal: means that the file contains a small nubmer of rows (count <10% of all rows in `ADNIMERGE`), and its impact can be minimal.
- Redundant: means that the information has been described by other files.
- Exclude: means that the file was mainly about one of the excluded cohorts, ADNI-1 or ADNI-GO, so not of interest for this study.

Outcome The protocols have changed several times on the encoding of the diagnosis. This has created a confusion in the usage of the terms in the ADNI diagnosis-based papers. For example there are papers that are basing their modeling on the `DX_bl` which is the screening diagnosis - different from the baseline diagnosis. The real value of the diagnosis for each visit is the `DX` variable. As seen below, the missingness of the diagnosis in visits was a problem widespread in the dataset. To not include further biases in the dataset, the visits with missing `DX` were dropped. This was done after observing that the diagnostic value from these patients could be unstable, with patients having transitions like: `HC-MCI-missing-MCI-HC-MCI-MCI`, so a replacement of these values would be creating bias.

3.1.2. National Alzheimer's Coordinating Center

NACC contains longitudinal multisite clinical data from patients of 37 Alzheimer's Disease Research Centers funded by the National Institute of Aging (NIA). More information can be found at their official website (https://naccdata.org/). A request that contains the intent of this research was sent for the permission of data usage. National Alzheimer's Coordinating Center (NACC) dataset has a more comprehensive inclusion of the neuropsychiatric and neuropsychological tests that are used clinically, and for some of the tests they have scores of results up to the question granularity.

The whole dataset is made of a total of 166082 visits of 45100 subjects. The version of the dataset is `investigator_nacc57.csv`. All the data is found in this file, and the information about the predictors is found in the accompanying guidelines. The subjects are diagnosed through protocols previously defined and it is not concentrated only on the HC-MCI-AD diagnosis evaluation. The different diagnoses are mentioned at section 2.3. Besides Dementia diagnosis, differential diagnosis subjects are included, like Pseudo Dementia because of psychiatric diseases. The predictors are extensive and inclusive, and the data from clinical examination takes precedence. Also, the information is stored closely to the descriptions mentioned in section 2.1. The data includes demographic information, family history of cognitive impairment, medical history including the medications, physical examination including measurement of factors like tremors and sleep disturbances (important for LBD), and others. Important are the explicit diagnosis protocols mentioned in the diagnosis section, with mentioning of the predictors that should be more important for the process.

3.1.3. Inclusion/Exclusion criteria

The subjects of these datasets are related mostly to a secondary care unit, after a referral. The number of centers included in the study centers is mentioned above. The main exclusion criteria were:

1. younger than 35: the age 35 was selected as the cutoff age because it has been observed to that the

patients with the early onset disorders of the Parkinson type show their symptoms earliest at this time,

2. missing npa or nps assessment: the exclusion of patients with no npa or nps were excluded because nps and npa are central to the clinical process and the research questions of this work are related to them.
3. diagnosis with other diseases than the Dementia grouping described in section 2.3. or healthy controls, or missing diagnosis: About the diagnostic outcome, having disorders that do not fit the Dementia groups, like other medical conditions or other Dementias like Huntington's Disease, would add more noise because of the lack of patients supporting a possible data-oriented diagnostic power.

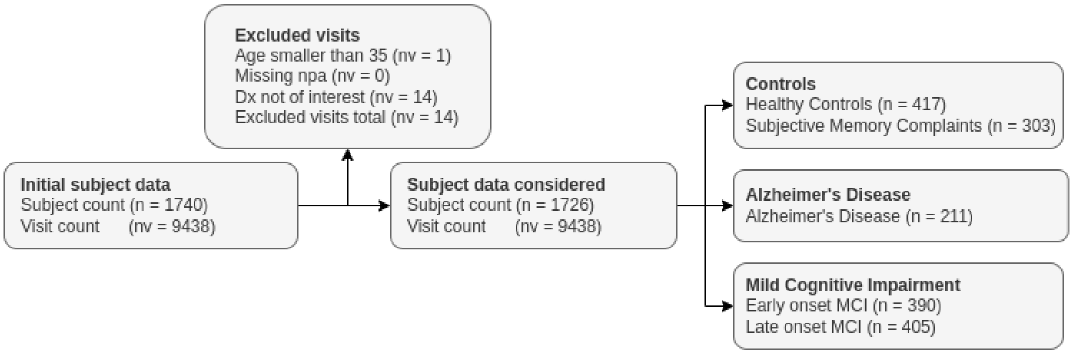
Excluded visits
Age smaller than 35 (nv = 1)
Missing npa (nv = 0)
Dx not of interest (nv = 14)
Excluded visits total (nv = 14)
Initial subject data
Subject count (n = 1740)
Visit count (nv = 9438)
Subject data considered
Subject count (n = 1726)
Visit count (nv = 9438)
Controls
Healthy Controls (n = 417)
Subjective Memory Complaints (n = 303)
Alzheimer's Disease
Alzheimer's Disease (n = 211)
Mild Cognitive Impairment
Early onset MCI (n = 390)
Late onset MCI (n = 405)

Figure 2: ADNI Participant Flow

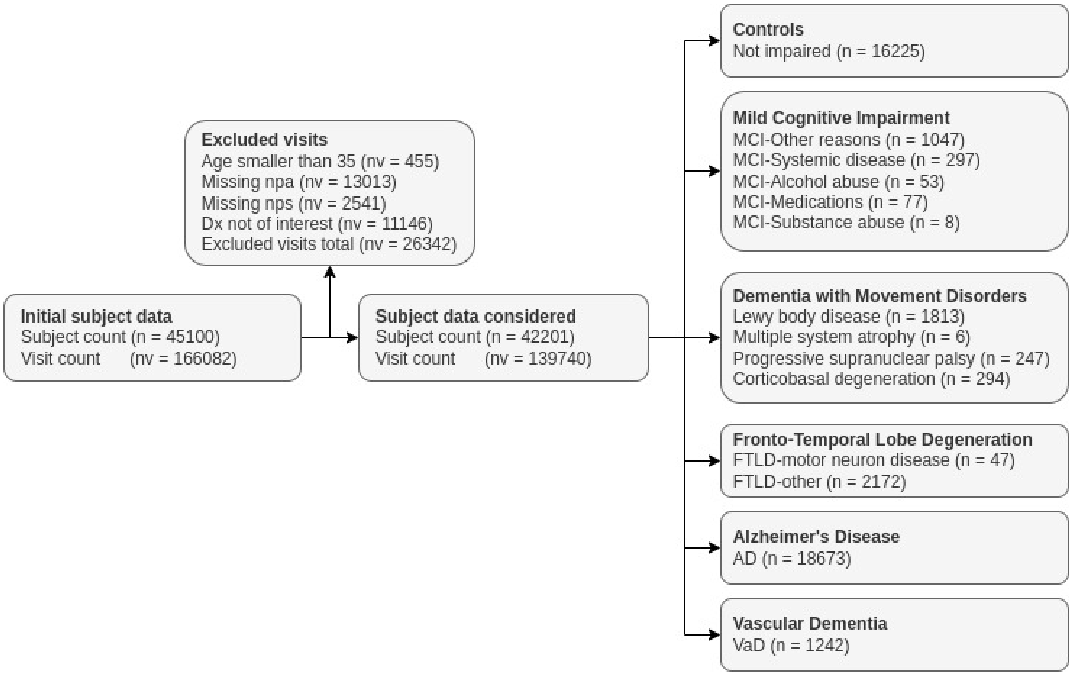
Excluded visits
Age smaller than 35 (nv = 455)
Missing npa (nv = 13013)
Missing nps (nv = 2541)
Dx not of interest (nv = 11146)
Excluded visits total (nv = 26342)
Initial subject data
Subject count (n = 45100)
Visit count (nv = 166082)
Subject data considered
Subject count (n = 42201)
Visit count (nv = 139740)
Controls
Not impaired (n = 16225)
Mild Cognitive Impairment
MCI-Other reasons (n = 1047)
MCI-Systemic disease (n = 297)
MCI-Alcohol abuse (n = 53)
MCI-Medications (n = 77)
MCI-Substance abuse (n = 8)
Dementia with Movement Disorders
Lewy body disease (n = 1813)
Multiple system atrophy (n = 6)
Progressive supranuclear palsy (n = 247)
Corticobasal degeneration (n = 294)
Fronto-Temporal Lobe Degeneration
FTLD-motor neuron disease (n = 47)
FTLD-other (n = 2172)
Alzheimer's Disease
AD (n = 18673)
Vascular Dementia
VaD (n = 1242)

Figure 3: NACC Participant Flow

3.1.4. Improbable diagnosis

Interestingly, the datasets in their diagnostification process suggest that the doctors make independent diagnosis from previous visits. The impacts of such process can be observed in the distribution of changes in the diagnosis for one patient. Let's define a improbable diagnosis as following: take a disease diagnosis from the set S={A, B, C, . . . } for a single patient, X can be any of S elements, and a patient's visits can be X1-X2-. . . -Xn, an improbable diagnosis is of the form X1-X1-X2-X1-X3 where X3 != X2, and X2 != `MCI`, as transitions to and out of `MCI` are possible, and X3 != `AD` as it might make sense that a patient with a different type of dementia eventually labelled as `AD`. For example, we can have patients that are of the following form: `HC-HC-AD-HC-HC`. The diagnostification as `AD` in this case is improbable. This kind of diagnostification skews the models, and eventually impacts the correct features.

Such definition can be handy when understanding what combination of factors can push a doctor into making a diagnosis that is not supported from the previous or following visits. This kind of 'errors' can be very valuable in the definition and understanding of limitations of clinical protocols. Following a generalized protocol can lead to such diagnosis that can be more of a data-problem than a clinical problem. Additionally, it can suggest some kind of over-generalization of the clinical classes that does not provide a degree of impairment. This over-generalization of patients, for example as `AD`, can impact the patient's life and also allows little space for possible treatments.

3.2. Statistical analysis methods

Step by step method

3.2.1. Preprocessing

The extensive amount of data gathered in the open datasets imposes constraints on the type of analysis that can be done. The data can be either numerical or categorical (includes binary). This separation allows for defining standard processing methods for each. A simple script for defining whether a feature is binary, categorical or numerical is attached at Appendix C. Binary features are treated as categorical from now on. The separation can be observed at Table 4 where besides the number of predictors, the numerical predictor count is also mentioned.

The preprocessing steps have been defined through standard processes tested in the literature. Still, some of these preprocessing methods are based on some assumptions of the dataset. In Appendix A the necessary preprocessing methods have been described in detail with the reasoning and descriptive statistics. Here a quick overview is provided.

- *Imbalance handling*: Is to reduce the high imbalance that might be either because of prevalence difference or because of some bias in the selection process. There are standard ways to do it through under-sampling or over-sampling, or a combination of both. Additionally, a `CustomUnderSampler` was created based on the high prevalence of the subjects that are constantly in a single state, like healthy controls or patients with Alzheimer's Disease.
- *Imputation*: Aims at handling the missingness represented in the dataset, and should differentiate between missingness subtypes. For example, missingness in a language task might come because the patient has behavioral problems, or maybe because they are observably good at it. Several strategies were tested: no imputation, `SimpleImputer`, no `IterativeImputer`.
- *Transformations*: Some of the features require some kind of handling for preparing them for the models. For example, the categorical data in ADNI dataset is in string form, but most classifiers accept only number format. So, encoding was necessary, either through `OneHotEncoding` or `OrdinalEncoding` (where allowed). For the numerical features potentially scaling can be a factor that can impact the results, so several types of scaling were tested.
- *Feature selection*: As it can be seen in the table below, the number of features is high (>1000) for both datasets. Adding these features in the models increases computationally the running time and distributes their importance. To reduce the feature space automatic feature selection processes were tested that either aimed at all features (through a simple initial model), or separately to numerical (variance or correlation) or categorical (Chi square based) features.

Predictor types The clinical research building the datasets follow different protocols, with several types of predictors. As shown in the table below, the most represented data sources are the neuropsychological assessment (npa), the neuroimaging data (nimg), and the neuropsychiatric assessment (nps). There is a difference in the types of predictors, with NACC dataset having more additional information stored that might be regarding the process of information, while ADNI has more extensive neuroimaging and neuropsychological assessment data. While the numbers are very high, most of these predictors suffer from a very high missingness, so they are dropped. For more details please check Appendix A.

Table 2: Study characteristics, and predictors (numerical).

Dataset	ADNI predictors	NACC predictors
Data collection period	2006-2021	2005-2022
Study design	Prospective cohort	Prospective cohort
Protocol	ADNI Protocol	UDS-3 Protocol

Dataset	ADNI predictors	NACC predictors
Outcome	HC-MCI-AD dx	Dementia dx
additional	0	226 (11)
co-participant	0	22 (3)
demographics	27 (13)	51 (17)
diagnosis	27 (17)	120 (0)
family-history	0	3 (0)
genetics	1 (0)	18 (0)
health-history	0	108 (8)
impair	74 (73)	20 (1)
medications	0	62 (1)
nimg	390 (365)	34 (1)
npa	347 (310)	133 (65)
nps	231 (227)	43 (1)
physical	16 (12)	125 (6)
text	0	59 (8)[2]
Predictor count	1113 (1017)	1024 (122)

3.2.2. Diagnostic models

Based on the literature, the models with the better results for heterogeneous data are ensemble models, with the multi-class outcomes and the diagnosis defined in each dataset. In the past decades the model has proven itself quite useful in medical and neuropsychological research including big datasets. Features like: parallel processing, simplicity, ability to analyze nonlinear-correlated data, preselection, and classification make the model indispensable. The models trained were:

1. `RandomForest`: RF was used as a baseline model. It requires the data to be complete, so it makes the preprocessing step of imputation necessary. It does not accept categorical data, so a process of encoding is necessary.
2. `XGBoost` (Extreme Gradient Boosting) is widely prefered among researchers as of its high predictive capacities. Its most valuable features are high efficiency in scenarios of regression and classification. It can handle natively the missing values, so it does not need imputation. It does not provide native support for the categorical features.

[2] While not in the form of text, an evaluation is taken at ‘B9 Clinician Judgment of Symptoms’ section.

3. `LightGBM` (Light Gradient Boosting Method, also known as `LGBM`) is a similar method to XGBoost, but more efficient and with better accuracy scores. In terms of explainability it is similar, but building the models is faster. Additionally it supports the encoded categorical features inherently (so, it does not require `OneHotEncoding`).
4. `DPEBM` (Differentially Private Explainable Boosting Machine) maintains the boosting capacities of the advanced models, but add a layer of privacy to the interaction (Nori et al. 2021).

Table 3: Preprocessing requirements for different models Logistic regression (LR) provided for comparison.

Requirement	LR	RF	XGB	LGBM	DPEBM
Numerical imputation	Yes	Yes	No	No	No
Numerical scaling	Yes	No	No	No	No
Categorical imputation	Yes	Yes	No	No	No
Categorical encoding	Yes	Yes	Partial	Partial	Partial
Private	No	No	No	No	Yes

3.2.3. Model performance

The models try to learn the underlying statistical distribution of the data producing process. The learning process can be done through the several ways described above. Still, there are stable comparative measures to compare and choose the algorithms that do better at this task. Some performance measures are overly-optimistic, while some others are more specifically interesting for certain types of classification.

- `Prevalence`: It provides the proportion of people that have the disease in the population measured in the dataset.
- `TP`, `TN`, `FP`, `FN`: defining how the model does in terms of real value and predicted value. *True Positives* (TP) and *True Negatives* (TN) are the values that the model found correctly, and *False Positives* (FP) and *False Negatives* (FN) are misclassifications.
- `Accuracy`: While not the main metric, accuracy can still be a good starting point to see how a model is doing. With imbalanced datasets, like the ones about Dementia, the accuracy does not provide good insight because the classes with more representation count more.
- `Sensitivity` and `Specificity`: Sensitivity shows the probability that the model correctly predicts disease when the disease is present, and Specificity is the probability that the model correctly predicts the lack of a problem when there is a lack of disease. Both of these metrics are not dependent on the

prevalence. Sensitivity is also known as `Recall` and specificity as `Precision`.

- `F1 Score`: It is the harmonic mean of Precision and Recall, and it has the highest value at 1 where both precision and recall are at their highest, and the lowest at 0. It can be used with cross-validation for having a balanced model.
- `PPV` and `NPV`: *Positive Predictive Value* (PPV) shows the probability that a subject predicted as having the disease, does indeed have the disease, expressed differently. Similarly, a *Negative Predictive Value* (NPV) shows the probability that a subject predicted as not having the disease, does not have the disease.

$prevalence = \frac{1}{N}\sum_i y_i$ where $y_i = 1$ where the patient has the disease.

$accuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$

$sensitivity = recall = \frac{\text{TP}}{\text{TP+FN}}$

$specificity = precision = \frac{\text{TN}}{\text{TN+FP}}$

$F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN}$

$PPV = \frac{\text{TP}}{\text{TP+FP}}$ and $NPV = \frac{\text{TN}}{\text{TN+FN}}$

For the models of diagnostic criteria, the `Sensitivity` measure is important because it reflects how well the models are able to `sense` the existence of certain disease, given that a subject has it. Still, if the interventive measures have a direct impact on the subject's life, the `Specificity` becomes also important. `F1` score is the integration of such measurements, and it can be used for picking the best model for clinical diagnstic methods. Below is a description of the common measurements more in detail. Additionally, for more custom needs (like giving more importance to the sensitivity of the class that is less represented) new scorers can be built. Another option was to use the scorer known as `F-beta Score`, with a `beta > 1` that gives more importance to the sensitivity of the models. In multiclass classifications each class contributes to the score based on their weight defined through their prevalence.

Visualizations

- `CM`: *Confusion matrix* is a matrix showing how the model behaves for each of the classes. It allows us to see which classes are mistaken mostly between each other, and can allow some insight for feature engineering for better classification.
- `PRC`: *Precision Recall Curve* shows the trade-off between these two metrics. An high area below PRC shows that the model is doing well for the prediction.

- `AUC`: *Area Under ROC* curve is a measure of goodness of fit. The *Receiver Operating Characteristic* (ROC) curve is the ratio between *True Positive Rate* (TPR) and *False Positive Rate* (FPR) where the ideal point is closest for `TPR ~ 1` and `FPR ~ 0` meaning that the predictive power is high, and the error rate is very low. The AUC, the area under this curve, represents how the model behaves under different thresholds. In the medical sense, AUC provides the intuition that a patient that has the diagnosis has a higher score than a patient who does not have it.

Validation There are several problems with developing a model that need to be handled carefully. Overfitting happens when the model learns the data very well, but it does not have predictive power. If a new case with a different profile from the previous data will be presented the prediction will be mistaken. This can be handled by splitting the dataset into training and testing by making sure that the test and training dataset do not share information. A more advanced method is the `k-fold` method where several models are trained with subsets of the dataset, and their predictive power merged.[3] These methods make sure there is no data-leaking. For datasets like the Dementia datasets, grouping the data of a single patient is needed. This makes sure that the same patient data is not represented both in the training and test process. Otherwise, the model learns that patient-X has a certain disease, and recalls it when tested, without actually learning. Additionally, there is a need to keep the ratio of the diagnostic classes equally distributed between the train and test, which is handled from the stratification. So, eventually `StratifiedGroupKFold` is used.

StratifiedGroupKFold for ADNI dataset.

3.2.4. Machine learning pipelines

As suggested from the section on 3.2.2, the diagnostic models have different necessary requirements, so coming up with a simple pipeline that does not satisfy those requirements would end in an error. So two vanilla pipelines were created: Vanilla-RandomForest and Vanilla-LGBM. For the ADNI dataset, the encoding of the categorical features was also added, while for the NACC dataset no further steps were necessary. Figure below shows the Vanilla pipelines for ADNI.

Facing with the complexity of the data and the possible necessary steps, the more complete pipeline having the following steps was built. Most of the steps are optional, and if not added the property "passthrough" of the pipeline can be passed.

- Imbalance handling (over- and under-sampling)
- Preprocessing

[3] Even though cross-validation through `KFold` is used to good extent by the community, it does not mean that it can be robust from the sources of variance (Bengio and Grandvalet 2004).

  - Categorical feature handling (Imputation, Encoding, Selection)
  - Numerical feature handling (Imputation, Scaling, Selection)
- To dense (sparse to dense block)
- Feature selection
- Classifier

Preprocessing is put together through the different transformers for both types of predictors (numerical and categorical). The last feature-selection step should be used if the previous ones have not been used. The to-dense transformer is for when the dataset becomes sparse in case of using one-hot-encoding with a large number of categorical features. The tuning of the hyperparameters was done through `GridSearchCV` on the pipeline for each process, and each step has been replacable by similar transformers. Additionally, future development is possible through replacing the grid-tuning through hypertuning.

Table 4: Transformers and estimators used in the pipelines.

Subprocess	Possible options
Imbalance	
Over-sampling	No over-sampler, `RandomOverSampler()`, `SMOTENC()`
Under-sampling	No under-sampler, `CustomHandler()`, `RandomUnderSampler()`, `NearMiss()`
Preprocessing	*categorical*
Missing data	No imputation, `SimpleImputer(*)`
Transforming data	No encoding, `OneHotEncoder()`, `OrdinalEncoder()`
Feature selection	No selection, `SelectPercentile(Chi2)`
Preprocessing	*numerical*
Missing data	No imputation, `SimpleImputer(*)`, `KNNImputer()`, `IterativeImputer(*)`
Transforming data	No scaling, `StandardScaler()`, `RobustScaler()`
Feature selection	No selection, `VarianceThreshold`, `SelectPercentile(Pearson)`
Method	
Classifier	`RandomForest()`, `XGBoost()`, `LightGBM()`, `DPEBM()`
Validation	
Cross-Validation	`StratifiedGroupKFold()`
Hypertuning	`GridSearchCV()`
Scoring	`f1_score()`, `fbeta_score()`

Vanilla2 Vanilla Not Vanilla

Figure 4: ADNI Pipelines - Vanilla RandomForest, (b) Vanilla LGBMClassifier, (c) Sample Grid-SearchCV.

3.3. Testing hypotheses

Firstly, the pipeline of processing needed to be validated, so extensive testing was done for its usability for different datasets. This could provide some comparison on how well was the processing method doing in comparison to the literature (Model1 is a replication of the latest work on Dementia modeling). Additionally, it could provide some comparison on the value added from different types of datasets (the predictors of AD and MCI found in Model2 could be compared with predictors found in Model3).

Table 5: Models built. (repr: representation)

Model	Dataset	Selection[4]	Disease (stable repr:dataset repr)
Model1	ADNI	Clinical:True, Category:basics, Source:ADNIMERGE.csv	HC:CN, MCI:MCI, AD:Dementia
Model2	ADNI	Clinical:True, Category:basics	HC:CN, MCI:MCI, AD:Dementia
Model3	NACC	Clinical:True, Category:basics	HC:88, MCI:30, AD:1, LBD:2, FTD:7, VaD:8
Model4	NACC	Clinical:True, Category:basics	HC:88, MCI:25-30, AD:1, DwMD:2-5, FTLD:6-7, VaD:8

(H1) Practical usefulness of minimal versus extensive assessment: the comparison will be done between Model1 and Model2 for seeing how a more extensive data gathering process can impact the model's capacity in predicting. Model1 is based mostly on the features mentioned in the literature, while the more extensive Model2 is based on the dataset merged.

(H2) The similarity of computationally selected features to the clinical protocols: A feature importance analysis using explainable machine learning was applied on each of these models, and the resulting feature importances and combinations were assessed in the light of diagnostic protocols. This analysis was based on Model3 and Model4.

(H3) The value of counterfactuals in providing possible intervention strategies for mis-classified patients: The miss-classified patients were analyzed using counterfactuals for understanding the minimal steps to change the diagnosis to the real class. This analysis was based on Model3 and Model4.

[4] `Clinical:True` means a first reduction based on redundancy or lack of added value. `Category:basics` is the exclusion of predictors that are part of the categories: `additional,text,medications,co-participant`, for reducing the predictor space, and concentrating on the relevant features. Please check the variable files to observe these selections.

Bengio, Yoshua, and Yves Grandvalet. 2004. “No Unbiased Estimator of the Variance of K-Fold Cross-Validation.” *J. Mach. Learn. Res.* 5 (December): 1089–1105.

Kumar, Sayantan, Inez Oh, Suzanne Schindler, Albert M Lai, Philip R O Payne, and Aditi Gupta. 2021. “Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review.” *JAMIA Open* 4 (3): ooab052.

Nori, Harsha, Rich Caruana, Zhiqi Bu, Judy Hanwen Shen, and Janardhan Kulkarni. 2021. “Accuracy, Interpretability, and Differential Privacy via Explainable Boosting.” In *Proceedings of the 38th International Conference on Machine Learning*, 139:8227–37. Proceedings of Machine Learning Research. PMLR.