University of Pavia
University School for Advanced Studies

Masters Degree in Psychology, Neurosciences and Human Sciences
Department of Brain and Behavioral Sciences

Revising the clinical criteria for Dementia using explainable machine learning.

Thesis of

Doren Çalliku

Supervisors:

Dr. Gerardo Salvato .....
Prof. Gabriella Bottini .....

Candidate:

Name Surname .....

Session of Graduation 30 September 2022
Academic Year 2021/2022

I, Doren Çalliku, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Background: Dementias are disorders of the brain that require an early intervention for slowing down the progression of the disease. Publicly open datasets have made possible the research for early markers for the various Dementia phenotypes. A research area of these markers is based on the machine learning models of diagnosis of Dementia.

Objective: This thesis uses the explainability analysis of these models to investigate whether similar criteria are shared between the clinical criteria of Dementia diagnosis and the machine learning criteria. This analysis can help specify the protocols and increase the sensitivity of the diagnostic process.

Design: First the clinical criteria for the main phenotypes of Dementia were defined. Then a pipeline that handles imputation, scaling, imbalance, and model optimization was set. The datasets were tested using different machine learning models like LightGBM, RandomForest, or LogisticRegression.

Setting: Two multi-center open datasets were mainly used for development and internal validation of the models: the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and National Alzheimer's Coordinating Center (NAC C) dataset. These two datasets include multiple visits for heterogenous groups of Dementia patients. They contain the data extracted during the clinical interview, questionnaires and assessments, and the diagnostic process. In one case (ADNI) processed neuroimaging data was also included.

Participants: There are 45100 participants in the NAC C dataset, and 2294 participants in the ADNI dataset.

Results: The better performing pipelines do not need additional preprocessing steps. An extensive non-neuroimaging evaluation of a subject is as good as a standard evaluation containing neuroimaging data. The clinical predictors described in the protocols can be improved and specified using the explainability of machine learning models.

Conclusions: Explainable machine learning models on clinical data can help shape the clinical criteria for the diagnosis of Dementia.

Acknowledgements

I thank my supervisors Gerardo Salvato and Gabriella Bottini, and my collaborators at Niguarda Hospital, Stefania Basilico and Maura Simioni. With their insights and their contribution, they made the thesis possible. I thank my sisters, Klerisa and Gresa, for their unbent support of my interests and my direction. My parents, Shpresa and Agron, have trusted and contributed to my work with their folk knowledge of neuropsychology. This path would have been very lonely without you. I am very grateful to the community working on Dementia, that has made the resources reachable, with an obsessive level of detail. To the open-source community of machine learning and the ones creating the tools that have been very useful to me since day one, thank you. To `mor ve ötesi`, Sapolsky, and Jahangirov, thank you for inspiring me.

Contents

List of Figures

List of Tables

Abbreviations

Symbol	Definition
Ab	**A**myloid-**b**eta
AD	**A**lzheimer's **D**isease
AxD	**Ax**ial **D**iffusivity
ADNI	**A**lzheimer's **D**isease **N**euroimaging **I**nitiative
APOE	**Apo**lipoprotein - **E**
APP	**A**myloid **P**recursor **P**rotein
CC	**C**orpus **C**allosum
CDR	**C**linical **D**ementia **R**ating
CSF	**C**erebrospinal **F**luid
fMRI	**F**unctional **M**agnetic **R**esonance **I**maging
FTD	**F**rontotemporal **D**ementia
GDS	**G**eriatric **D**epression **S**cale
ICV	**I**ntracranial **V**olume
ID	**I**mpulse **D**yscontrol Domain
MCI	**M**ild **C**ognitive **I**mpairment
ML	**M**achine **L**earning
MRI	**M**agnetic **R**esonance **I**maging
NACC	**N**ational **A**lzheimer's **C**oordinating **C**enter
NC	**N**ormal **C**ognition
NPI	**N**euro**p**sychiatric **I**nventory
NPA	**N**euro**p**sychological **A**ssessment
ROC	**R**eceiver **O**perating **C**haracteristic
SCD	**S**ubjective **C**ognitive **D**ecline

Chapter 1. Introduction

The pathological aging processes that are reflected in a drop in cognitive abilities are defined as diseases of Dementia. Dementias affect over 47 million people worldwide (Organisation 2017). In developed countries, with an increase in living standards and an increase in age, the number of older adults with Dementia is increasing. More than 1.3 million demented people lived in Italy by 2019, and the predicted prevalence by 2050 is expected to reach more than 2.2 million (Alzheimer Europe 2019). Developing countries are facing bigger public health challenges because of Dementia. For example, Turkey and Brazil had the highest age-standardized prevalence as of 2016, with more than a thousand individuals affected for every a hundred thousand (Nichols et al. 2019). This increases the need for easy to use tools that help with the clinical management of the disease until a cure has been found. For the various forms of neurodegenerative diseases, existing literature has pointed out a series of biomarkers that can distinguish between them and that are prognostic of their evolution. They comprise the presence genetic mutations (e.g., the MAPT mutation in Cortico-Basal-Degeneration, (Kouri et al. 2014)) or specific brain pathologic changes (e.g., the presence of hippocampal atrophy in Alzheimer's Disease, (Pini et al. 2016)). Over the years evidence has mounted supporting the diagnostic and prognostic role of patients' performance in neuropsychiatric or neuropsychological tests, which could be used as useful early markers in the diagnostic process. Some of the important differences are captured only in the distribution of errors in neuropsychological assessment (NPA) (Salmon et al. 2002). For example, differentiation between Alzheimer's Disease (AD) and FrontoTemporal Lobe Degeneration (FTLD) patients is possible through a qualitative analysis of visuospatial and visuo-constructive deficits measured by their corresponding tests. Many of these qualitative differences have been proven through subsequent autopsy-confirmed data (Salmon and Bondi 2009). The search for the neurological (brain) markers for neurodegenerative diseases like Alzheimer's has produced a large bulk of computational models with an inclusion of demographic, minimal NPA, neurological, and neuroimaging data. The question most of these models ask is: will a patient in a Mild Cognitive Impairment (MCI) stage progress into Alzheimer's Disease? The results have been promising, with high accuracy of predicting progression for several models (Kumar et al. 2021; Tosi et al. 2020). The development of computational models holds the promise that one day a simple clinical examination combined with a neuroimaging procedure will be able to accurately predict the route of a disease in an individual patient. The benefits of such a large field of research have allowed for a general scrutiny that has improved the data, practices and models related to the study of Dementia. The latest models as of now are the explainable machine learning models built on the ADNI dataset that can predict a transition to AD from MCI accurately, and the decisions can be explained using the features (Bogdanovic, Eftimov, and Simjanoska 2022, @Hernandez2022).

1.1. Problem description

While the models of diagnosis for Dementia have been developing, several problems have arised related to the cost, specificity, and ability to personalize them.

Most of the research is concentrated on neuroimaging data because the power of machine learning has been observed mostly in the Convolutional Neural Networks direction. This intuitive direction for looking for neuroimaging biomarkers requires the clinical practice that is not cost-efficient and is based on well-funded initiatives. A more integrative approach needs to be taken, concentrating more on the cheaper methods like the clinical data provided by the other methods of assessment.

Based on the prevalence of Alzheimer's Disease, a good proportion of the models have been built to predict the progression in the spectrum of Healthy-MCI-AD (Kumar et al. 2021). This significant amount of research is closely related also to the availability of the datasets and the previous literature that makes them easier to compare. The problem is that the cases in a clinical setting can be more complex, with patients showing symptoms of FTLD and Parkinsonisms. In this context the models described have no power because they have been tuned to be sensitive to one subgroup of Dementia. Having models that consider a wide range of possible diagnose is necessary.

The success of these models is based on the integration level to the medical practice. Still, the research of the field deals mostly with the problems through its engineering problem: optimizing a model to improve metrics like precision or f1-score. This frame of thinking in terms of benchmarking diminishes the significance of the models for the clinician. For the bulk of the cases that the models predict well, the doctors do so too; for the ones that the model does better the explanation tends to be in the non-clinical language. Having better explainable models that integrate clinical knowledge with computational methods is essential to the future usage of these tools for the diagnosis of Dementia.

1.2. Research questions

This thesis aims to further develop further and validate existing models of Dementia diagnosis by improving the pipeline of analysis. The research questions this thesis aims to answer:

1. What is the variety of computational models that can be created, and how to select the better ones for diagnosis of Dementia?
2. Are automatically selected predictors of the computational models similar to the predictors defined in the protocols of diagnosis of Dementia?
3. Can computational models built on an extensive clinical assessment (like the NACC dataset) can have a

similar predictive power as the models built using neuroimaging data and a restricted group of clincial data (like the ADNI dataset)? While the neuroimaging data presence is not central to the thesis, it is important to compare the new models with the previous models of the literature.

4. Can these models provide possible improvements that can be done for specific cases by checking the best route to changing a diagnosis?

1.3. Significance

There is a large research area on Dementia biomarkers, mainly concentrating on the performance metric of classification. To the knowledge of the author, this is the first work to:

- use a detailed reproducible pipeline of analysis for the NACC dataset, an extensive Dementia dataset,
- test how the clinical criteria for Dementia are represented in the models for diagnosis in diseases like Frontotemporal Dementia and Lewy Bodies Dementia,
- provide a metric of instability of the diagnosis with possible routes through counterfactuals.

Additionally, in this work can be found detailed discussions on Feature Selection, Missing data, and Dataset Imbalance that can be useful for future directions for improving the diagnostic models of Dementia.

1.4. Organization of chapters

The thesis is organized into six chapters. In the first chapter "Introduction" the scale of the problem and its importance is presented, together with the related research questions, and the significance of this work. Additionally, the organization of the chapters is given. In the second chapter "Theoretical Background" a description of the clinical examination, the normal aging process and the possible neuropathologies is presented. Then the latest computational models with possible problems that these models face like predictor selection, missing data, and imbalanced classes are described. Then the explainability of these models is shortly mentioned in terms of Shapley values and counterfactual explanations. This complete overview serves to have a context for the hypotheses. In the third chapter "Methodology" the components of testing the hypothesis is described by introducing the datasets and the participants involved, the clinically derived features from the protocols, the machine learning model building process, and the explainability of such models. The possible limitations of these multi-step procedures are described and the attempt to systematically minimize them is explained. In the fourth chapter "Results" the results of the experimentation with the dataset and the method are provided. The findings are explained in the light of the research questions and the usage of the models for direct clinical insight is shown. In the fifth chapter, the "Discussions" the extent of implications of the results is mainly discussed in the context of personalizable diagnostic path. The limitations of the study

and the degree of interpretation are presented. In the sixth chapter, “Conclusions”, the results are presented in the light of the theory and recommendations for future research are provided.

Chapter 2. Theoretical background

In light of the latest developments in open-source data science tools and the aggregation of longitudinal data from the neuropsychological units, we can reproduce prior research and create new biomarkers with predictive value. The literature on the cognitive resilience of individuals with dementia suggests the prospect of helping these patients with different rehabilitation protocols. Cognitive resilience can improve a patient's daily life by enabling them to be independent and maintain their functional routine.

2.1. Normal aging

Aging challenges the brain to maintain cognitive abilities central to daily life. The aspects of independence and socialization, two central features of human nature, face major difficulties. Some of these ongoing processes are worrying and need attention, and for some elderly the process is more pathological. Defining the cause, the probable progression, and modification strategies can help improve the life of these old adults, and even more - provide an insight into the process of aging.

Chronological aging tends to be accompanied by cognitive changes that condition the life of older adults. The brain structure and activity patterns also change. An example is the bilateral activation compensatory mechanism that helps maintain a similar performance for tasks of perception. Some of the impacts of structural changes can be directly observed in the cognitive abilities of the elderly. The decrease of White Matter (WM) density affects the processing speed of most of the networks (**???**). This in turn makes the elderly slower at processing multiple inputs at once, so they can get confused from complex sensory inputs. There is a loss of dopamine receptors that affects the attentive capacities and concentration on goals (**???**). Neurofibrillary tangles (NFTs) and senile plaques (SPs) start appearing initially in the limbic region, damaging the areas that are responsible for processes like memory and emotion (**???**).

Neuropsychological theories of aging Several theories try to explain the relationship between the structural brain changes and cognitive aging. These theories deal mostly with the reorganization of the brain and re-purposing of networks like the Default Mode Network (DMN).

- *Dedifferentiation* is the reduction of neural specificity that happens in old age that forces some cognitive abilities to be performed through compensatory mechanisms (Persson et al. 2006). For example, learning in young adults activates primarily hippocampal regions, while its activity involves mainly frontal regions in old people.
- The *processing speed theory* suggests that loss of white matter forces slow processing. This increases resistance to the consumption of simultaneously complex information (Salthouse 1996). For example,

the attention system performs worse in old age because it is difficult for the brain to shift processes at the right *speed*.

- The *scaffolding theory* considers the brain as a responsive system that, in the face of adversity because of structural changes, shifts the processing of main tasks from specific regions to more generalized regions (Park and Reuter-Lorenz 2009). This keeps the cognitive performance stable but relies heavily on the networks of the frontal regions of the brain, and on networks like the DMN.

Cognitive reserve While the structural brain changes have been observed to impact cognition, the degree of change does not have predictive power for the future development of the aging process. Some people with largely structural brain damage can function properly, while some others with less damage might be more cognitively impaired. The high variability expressed has been related to the individual differences (health, education, occupational complexity) and contextual factors (living with a partner, location) (**???**). Cognitive reserve has been proposed as an internal factor of the brain's overall elasticity and connectivity that can explain to some level this mismatch between the level of impairment and the structural brain damage (Stern 2009). Particularly a healthy lifestyle that comprises physical activity and a rich social life seem to be protective factors (**???**). This lifestyle increases the protection from cardiovascular diseases that are closely related to mild impairment from vascular dementia. Also early life education and continuous occupational complexity seem to help in the formation of elegant connections between networks that can be resilient to damage (**???**).

2.2. Neuropathology of aging

In the following subsections the different neuropathologies related to Dementia will be mentioned. They have been grouped based on their symptoms and on the literature so far. For each group there is a table of predictors that shows the type of predict, the category based on section 2.1, and some notes on the differential power or some reference.

2.2.1. Mild Cognitive Impairment

Mild Cognitive Impairment (MCI) is often referred because of a subjective decline of cognitive abilities that interferes with the daily life of the patient. MCI patients' results are lower than elderly at their age, but their cognitive impairment is not at the level of people with Dementia (Petersen 2004). The earlier criteria were concentrated on the subjective memory complaint, the MMSE-score higher than 24 and the occupational impairment. The updated clinical criteria for MCI according to the International Working Group (IWG) on MCI (Winblad et al. 2004) are more generalized, and include impairment in one or more cognitive domains,

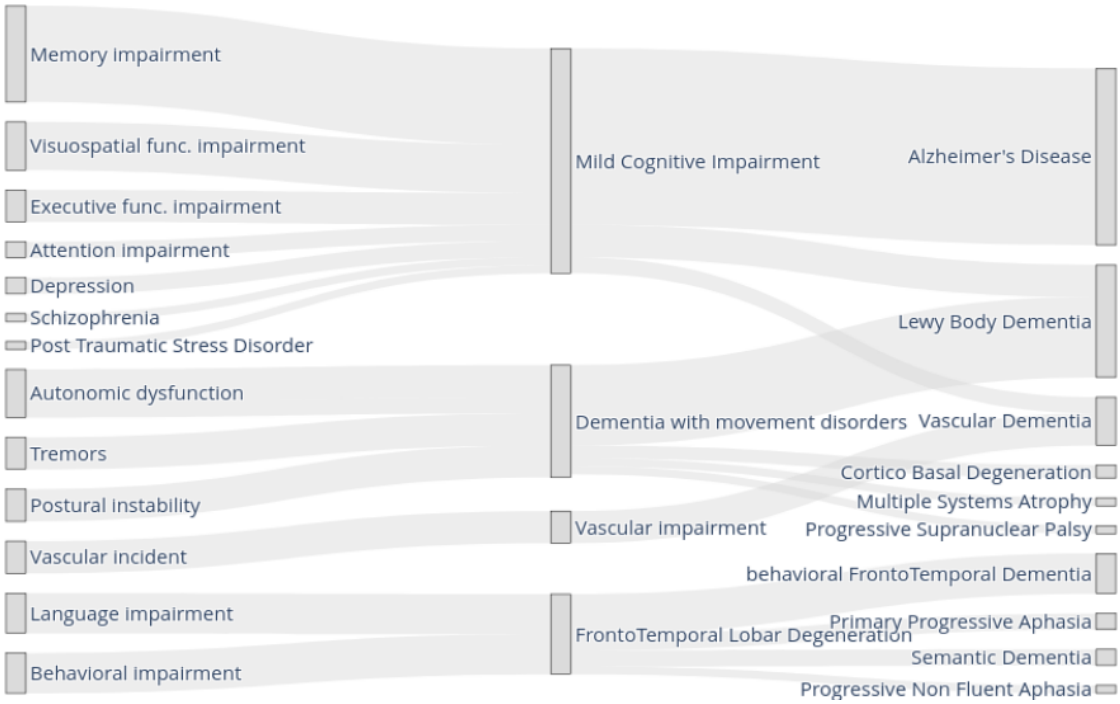

Figure 1: Transitions to dementia, a simple illustration.

and a level of preservation in functional abilities. The neural correlates of MCI tend to show a volume loss in the temporal lobe, the cingulate, and precuneus. Additionally, the ventricular (CSF) volume increases depending on the type of MCI. The neuropathology found through Pittsburgh compound B (PiB) using Positron Emission Tomography (PET) shows an increase of PiB retention for MCI patients. The degree of impairment of the different domains together with the etiology of the MCI can have prognostic importance. For example, patients with a-MCI tend to convert into Alzheimer's Disease (Albert, DeKosky, and Dickson 2011), while the ones with na-MCI in a domain like attention or visuospatial functions tend to convert to Lewy Bodies Dementia (McKeith, Ferman, and Thomas 2020). The MCI phenotypes can be separated based on the presence of memory problems (amnesic or not), number of domains mainly impaired (single- or multi-domain), or on the progression state (stable or progressive).

- *Amnestic (a-MCI)*: patients have predominantly memory problems.
- *Non-amnestic (na-MCI)*: patients are impaired in other areas of cognition, be them the executive-functions or visuospatial.
- *Single-domain (sd-MCI)*: patients tend to be mainly on the memory field, but not necessarily.
- *Multi-domain (md-MCI)*: patients impairment might impact memory and another function like attention or visuospatial functions.
- *Stable (s-MCI)*: patients tends to maintain a level of impairment.

- *Progressive (p-MCI)*: patients convert to some type of Dementia.

2.2.2. Alzheimer's Disease

The most common form of neurodegenerative disorder is Alzheimer's Disease (AD). Besides the general decline centered on amnesia, there are differentiating features of AD patients: they have abnormally rapid forgetting, less effective semantic encoding, intrusion errors, extended difficulties in concurrent manipulation of information, and more (Salmon and Bondi 2009). The genetic factor of APOE allele e4 can increase the chances of AD progression (**???**).

For AD, the related brain pathologic changes are the neuronal atrophy, synapse loss, amyloid protein forming of plaques, and creation of neurofibrillary tangles by tau mainly in the medial temporal lobe limbic structures, and then their distribution (Braak and Braak, 1991). There is a tendency of these biological factors to generally create a brain volume reduction, and specifically disconnect cortical layers (neocortex) and the limbic system (entorhinal cortex, hippocampus). Tau pathology first appears in the trans-entorhinal region, followed by the limbic and neocortical areas (**???**). These spatial patterns are correlated with the different clinical stages of the disease. With the progression of the disease, the severity of the symptoms increases, thus increasing the dependence of individuals on caregivers. Changes in personality, emotions, and social functioning are also evident in individuals with AD (**???**). Individuals that meet the core clinical criteria of AD demonstrate progressive cognitive and functional decline on neuropsychological assessments, with the additional biomarker evidence of amyloid-beta (Ab) deposition, and elevated CSF tau increasing the certainty of AD pathophysiological process. The clinical diagnosis is based on the 2011 NIA-AA criteria for IWG on AD (McKhann and Knopman 2011). The phenotypes of AD can be described as follows:

- *Early onset AD*: starts earlier than the other ADs with memory decline. Tends to have a quicker downturn.
- *Possible AD*: amnesic or non-amnesic AD with an atypical course or without evidence for a cognitive decline.
- *Probable AD*: Gradual onset with memory impairments that start from episodic memory, an in-ability to form or retain new memories, and expand to the other areas of cognition.
- *Prodromal AD*: patients with mild cognitive impairment but with high levels of pathology in the brain. It is observed at people with a high cognitive reserve.

2.2.3. Dementia with movement disorders

Parkinsonian signs and symptoms tend to be grouped as autonomic dysfunction, tremors, slow movement, muscle rigidity, and postural instability. Atypical parkinsonism is not mentioned. The main phenotypes are described below:

- *Lewy Body Dementia (DLB)*: characterized by cognitive impairment, parkinsonism motor symptoms, fluctuating mental status, visual hallucinations, and rapid eye movement (REM) sleep disorders. Clinical criteria diagnosis based on IWG on DLB (McKeith, Boeve, and Dickson 2017).
- *Corticobasal Degeneration (CBD)*: characterized by limb apraxia, rigidity, dystonia, movement deficits, and sensory loss (Armstrong et al. 2013). Mainly related to executive dysfunction and visuospatial functions.
- *Progressive Supranuclear Palsy (PSP)*: patients exhibit downgaze palsy, retropulsion, and parkinsonisms. The pseudo-psychiatric signs are uncontrollable laughter or crying. Cognitively they are dysfunctional in terms of executive functions and working memory. Clinical criteria according to the NNIPSS study (Brain 2009)
- *Multiple System Atrophy (MSA)*: patients exhibit cerebellar ataxia, impotence and vocal cord paralysis among others. Clinical criteria based on IWG on MSA (Gilman, Wenning, and Low 2008).
- *Amyotrophic Lateral Sclerosis (ALS)*: patients show a pattern of lower motor degeneration, and progressive spread of the symptoms.

2.2.4. Fronto-Temporal Lobe Degeneration

Frontal lobe degeneration is a syndrome term used for many progressive diseases that affect mainly the language or behavior of the patient. The early onset, together with the co-occurrence with motor neuron diseases, tend to separate them from the other dementias (Josephs 2007). The degeneration tends to be lateral in the temporal lobe, and a volume decrease tends to be significant. The clinical criteria vary for the phenotypes. These individuals typically have predominant frontal temporal atrophy on structural MRI, high levels of executive dysfunction and characteristic patterns of frontal and temporal lobe decreased metabolism or blood flow, and functional neuroimaging such as PET or SPECT. The phenotypes of FTLD can either be classified by their symptoms (bv-FTD, SD, PNFA), or by their causes (FTLD-TAU, FTLD-TDP, FUS-FTLD). We here describe only the first group, as it concentrates more on the clinical profile, and this is useful to our aim:

- *behavioural-variant Fronto-Temporal Dementia (bv-FTD)*: patients show sharp personality and behavior

change, with apathy and disinhibition being main drivers of impairment (Rascovsky et al. 2011). Co-occurrence with ALS.

- *Semantic Dementia (SD)*: patients have a fluent aphasia, where they talk but lose the semantic meaning. Left degeneration is observed than right degeneration. Tends to be related to TDP-43 caused FTLD.
- *Progressive Non-Fluent Aphasia (PNFA)*: patients show mainly a difficulty in speaking and agrammatism. Co-occurrence with CBD and PSP.

2.2.5. Vascular Cognitive Impairment

Imaging evidence for Cardio-Vascular Disease according to each center and the accepted practices (Wardlaw, Smith, and Biessels 2013). Cerebrovascular diseases might affect one's cognitive sphere and cause impairment of different severity. The several Vascular Cognitive Impairments are sub grouped into Vascular Dementia (VaD), Vascular Cognitive Impairment No Dementia (VCIND), Mild Cognitive Impairment (MCI) and more (Wardlaw, Smith, and Biessels 2013). Adhering patients to one group or another might be challenging for the clinicians as sometimes it is unclear if the impairment is caused or rather revealed by the Cerebrovascular event. Some way to distinguish them is through a 12-month follow up neuropsychological tests for attention, executive function, clock , psychomotor speed, activation/language, visuospatial, memory and Beck Depression (REF?). However some of the case examples e.g. those having VaD might differ themselves from others not only through tests but also through recognisable loss of several daily life functionalities (REF?).

- *Vascular Dementia (VaD)*: VaD is the syndrome with the most severe impairment from the ones mentioned above. It is caused by the lacunar infarcts, and/or by the presence of stroke in the cerebral arteries. Due to VaD the patient might have several cognitive deficits combined such as memory impairment, cognitive disturbances(executive functioning, aphasia, apraxia, and agnosia), physical symptoms(extrapyramidal, bilateral pyramidal, positive masseter reflex, imbalance, incontinence, dysarthria, and dysphagia), and depression. (REF?). A patient is diagnosed with VaD after they have gone through a stroke and/or infract, reckon the above mentioned symptomes and scores low in above mentioned tests.
- *Mild Cognitive Impairment (MCI)*: Relating MCI to cerebrovascular diseases is at times as difficult as separating it from Normal Aging. Similar struggles do clinicians have with VCIND. Neuropsychological tests data come handy yet when backed up with Neuroimaging evidence (e.g., 3T MRI, 1.5T MRI, FLAIR) they are more trustworthy and state defining (REF?).

2.3. Clinical data

Clinical interview The present state of the patient can be better assessed by understanding the disturbances in their cognition. Understanding the compensatory strategies that they are using, on the other hand, is as essential. The clinical interview can help the process of gathering and allows an assessment of the patient's behavior and responsiveness. It includes the severity and the progression of cognitive symptoms, their impact on daily life, the patient's awareness of their problem, attitude, mood, spontaneous speech, and behavior. Table 1 shows a summary of the information stored. Potential physical problems examination follows. This process helps select the breadth and depth of the questionnaires and assessments for the patient. It can be a starting point to create hypotheses for the possible neuropsychological correlates and set some predictions for the test results.

Demographics: Asking for demographic information is a convenient way to observe the basic level of insight the patient has. Information like age, gender, handedness, language, and who referred them are collected. Age, as mentioned, is a primary risk factor, with people having a high probability of Dementia once they are older than 80-85 years.

Symptoms: At the same time, the initial visit can be valuable to track when the symptoms evolve into impairment and help understand their nature. The symptoms can be physical, cognitive, or emotional. The "Physical assessment" subsection examines the physical symptoms. Cognition-wise, symptoms start mainly with memory complaints, but they might have underlying causes such as language or attention. Emotional symptoms are also of great importance, such as personality change, depression or anxiety, or psychosis. The proper evaluation and characterization of such symptoms can provide good differential diagnosis power to the clinician. Both cognitive and emotional symptoms are assessed further with questionnaires and neuropsychological testing.

Physical: The physical assessment can help reckon some physical difficulty that comes with age and impacts the test results. The physical symptoms can be motor, sensory, or corporal. For example, hearing loss can affect the results for most of the tests but does not imply that there is cognitive impairment. Still, some motor symptoms, such as tremors or an account of falls, can set possible safety concerns.

Health history and medications: Health history can be an indicator of risk factors for cognitive impairment. For example, having a stroke increases the risk of having Vascular Dementia. Some of the medications can directly impact the cognitive abilities of the subject. For instance, one's attention can be affected directly by neuropsychiatric drugs, which can impact motivational circuits.

Family history: Having a first-degree relative affected by cognitive impairment is significant to the diagnostic

process. Some genetic factors can increase susceptibility to cognitive impairment, for example, Huntington's Disease.

Caregiver presence: The partner, a family member, or a friend generally accompanies the subjects coming for testing. The co-participant can help with a secondary view of the patient's state through a parallel clinical interview and questionnaires. They can inform about the level of impairment, physical symptoms, neuropsychiatric issues, medications or additional issues on the subject. The level of detail is dependent on how close they are to the subject.

Additional interpretation (text): Common to the clinical reports are also the clinical notes. This discharge summary can have nuanced observations that are not included in the tests, and might be related to some behavior that is to be noted later.

Clinical examination After conducting the clinical interview, the patient goes through some questionnaires and testing defined by the examiner. When the extensive assessment is not enough to define the diagnosis, more advanced tests are required, like neuroimaging. The type of neuroimaging test administered is chosen based on the clinician's hypothesis. Table 2 depicts a summary of these questionnaires and tests.

Impairment (impair.): Functional capacity to handle their activities of daily living (ADL) is of primary importance to the patient and their family. A basic ADL(bathing, dressing, feeding, etc.) can assess the independence of the patient and the higher-order abilities by the instrumental ADL (pay bills, shop, meal, etc.).

Neuropsychiatric examination (nps): The emotional state of the patient can be a marker of the pathology of their brain. The application of NPS is traditionally conceptualized as non-cognitive symptoms and include impairments of mood, anxiety, drive, sleep, appetite, and behavioral disturbances such as agitation. A perceived loss of independence and an awareness of the demented patient's cognitive decline can induce depression or anxiety in more than 20% of the patients (**???**). For instance, in patients that present with depressive symptoms separating the Dementia because of depression (because of lack of social bonding) and depression because of Dementia can be challenging. The questionnaires need to be followed by an in-depth understanding of the context of the symptoms by interacting with the patient and their caregiver.

Neuropsychological assessment (npa): Neuropsychological assessment allows an in-depth and complete analysis of each patient's cognitive and emotional status. This assessment allows a comparative profiling of the patient with standard measures, controlled per age and socioeconomic background. The neuropsychologist assesses aspects of the human cognition to understand the ability and vulnerability of the patient. The

neuropsychological assessment can follow different protocols, and tests can be selectively administered to the patient depending on the hypothesis (created in the clinical interview) and the assessed ability to respond to long testing. Sample tests with their descriptions can be found at Appendix B.

Neuroimaging (nimg-)*: In difficult cases where the stage or the type of the disease are not clear, radiology based imaging processes can be used, such as Magnetic Resonance Imaging (MRI) or PET Positron Emission Tomography (PET). Additionally, for a more comprehensive list of neuropsychological tests for Dementias like AD and their power to explain the stage of the disease read the meta-analysis by Dukehan and colleagues (Duke Han et al. 2017).

Biomedical (biomed) and Genetics (genetics): Biomedical data like Cerebrospinal Fluid have proven valuable in distinguishing Alzheimer's Disease from other Dementias such asLewy Body Disease. Moreover, the APOE4 gene has been found as an important risk factor for developing AD. While the data related to these biomarkers are less prevalent in the open datasets, it is still important to acknowledge their diagnostic power.

Diagnosis (dx): After gathering most of the information, the clinician sets a possible or probable diagnosis, and defines the cognitive impairment level. It also defines what were the main reasons for the diagnosis, including the biomarker effect. This allows a future validation of the reasoning; and keeping track of patients affected - in case the clinical criteria change because of new research. The etiology (EX) of the disease does define a probable route for the patients, defined as prognosis. The differential diagnosis is the disease that shows a close phenotype, and for which the model of diagnosis needs to show care.

Table 1: The elements of clinical examination.

Category	Interest	Example Predictor
Clinical	**interview**	
demographics	demographics	age, race, gender, etc.
demographics	initial-visit	year
co-participant	co-demographics	relation, age, race, gender, etc.
physical	physical	others
physical	sensory	vision, hearing, taste, appetite, etc.
physical	motor	gait, weakness, tremor, stiffness, etc.
physical	somatic	sleep disturbance, headache, etc.
physical	sleep	REM sleep disturbance, etc.
physical	measurements	height, weight, body-mass
health-history	health-history	stroke, surgeries, substance use, etc.

Category	Interest	Example Predictor
health-history	cardio-metabolic	ischemia, type-2-diabetes, etc.
health-history	self-care	blood-sugar, diet, exercising, etc.
health-history	psychiatric	depression, anxiety, psychosis, etc.
health-history	neurodev	learning difficulties, autim, etc.
family-history	family-history	family-cogimp, parent-cogimp
medications	medications	x-medicine
medications	dementia	Donezepil (AD), antiparkinson, etc.
medications	compliance	forgetfulness, preference, etc.
text	#text	additional signs
Clinical	**examination**	
impair.	impairment	Basic-ADL, Instrumental-ADL
impair.	onset	early, late
impair.	progression	sudden, gradual
nps	general	NPI-q, FQI
nps	depression	Beck Depression Inventory II
npa	global	MMSE, MDRS, MoCA, ADAS-Cog, STMS, WMS, etc.
npa	memory	AVLT, BVRT, CVLT-II, Detail Acc., FRsrt, ISLT, etc.
npa	language	BNT, CAT, CF, FAS, Phonemic Fl., SM, Animals, etc.
npa	visuospatial	BPSO, GMLT, PDT, RCFT, Rey, VFDT, WAIS Block, etc.
npa	processing-speed	DSST, WAIS Digital, WMS-R Digit, Motor Speed, etc.
npa	executive-f.	TMPt. B, WCST, ToL, Stroop, Mazes, etc.
npa	attention	useful field of view, cancellation task, etc.
npa	intelligence	WAIS, etc.
nimg(vol)	#region/#extra	sMRI
nimg(act)	#region/#network	fMRI
nimg(pat)	#region	PET
nimg(fra)	#network	DTI
biomed	#biomed	CSF
genetics	#gene	APOE4
diagnosis	dx	Healthy, MCI, AD, VaD, FTLD, DLB, PsycD, etc.
diagnosis	prob.	possible, probable

2.4. Clinical protocols

The clinical protocols for Dementia are results of observations of the clinicians on the nature of the diseases, together with the biomarker findings. For most of these diseases an International Working Group (IWG) with specialists comes together and they define or revise existing clinical criteria. Some of these IWGs have been gathered for MCI (**???**), AD (**???**), DLB (**???**), or bv-FTD (**???**).

The clinical criteria are defined in terms of significant predictors, and existing biomarkers. With the development of new biomarkers, the clinical criteria change. An example of this kind of change has been seen in the AD clinical criteria: the biomarkers once a supportive feature, now are central to the diagnosis. These kinds of changes require training for most of the medical doctors, and they increase the need for the integration with the technologies for all the hospitals that deal with patients with Dementia. Once that the problem includes the subtypes of the diseases, and possible interactions with them, the system of protocols does not maintain its level of practical knowledge.

Definitions A simple description of these protocols contains many layers: predictor, observation, clinical criterion, protocol. Testing the significance of each of these layers remains difficult because of the complexity of the disorders and the lack of data. Still, a tentative of defining them looks like this:

- Predictor (`p`): The predictors can be either a core clinical feature, or a supportive feature. They can be symptoms (ex. memory measurements), process descriptors (ex. sudden onset), or biomarkers (ex. amyloid PET trace).
- Observed state (`o`): It is a level of impairment or lack of impairment that can be described through the score of the predictor `p`. A possible score can be the equivalence scores of tests that are in the range `[0-3]`, where `0` means deficient, while `3` means completely normal. An observed state would be: $p_1 > 1, p_2 = 3, p_3 = 0$ where the subject is deficient in p_3, but is okay for p_1 and p_2.
- Clinical criterion (`cc`): It is the unit of the clinical criteria. It can be either core or supportive, and either inclusive or exclusive. It is made of combination of predictors that can either be simple or conditional. Put in another way, let us define `L` as a list of observations that are described in the clinical criterion, o_i as predictor in `L` ($o_i \in L$), `len(L)` as the size (length) of the list, and `O` as the list of predictors observed in the subject ($o_i \in O$):
    1. **simple cc**: `x` observations out of `L` are present in `O`: $x <= len(O) <= len(L)$. This includes also the case of all predictors in list, where $x = a$. It also considers the cases that the observations might not be done all, as the tests might be lengthy or expensive.

2. **indirect cc**: contains some kind of count (ex. instances of falling in the last one year), or some kind of unmeasurable predictor.
3. **conditional cc**: the combination of multiple clinical criteria through `and`, `or`, `xor`, and `not`.

For example, one valid typical AD diagnosis is defined when the patient experiences a progressive change in memory, and there is some biomarker validation of the tracer retention on amyloid PET. Additionally, we know that if the patient shows some impairment in the movement disorders the etiology of the disease will include some kind of movement disorders. Similarly, if the changes in the personality are big in an early age then FTLD disorders are more probable. In this case:

- Predictors: $p_1 = memory, p_2 = nimg(pat), p_3 = movement, p_4 = behavior, p_5 = age$
- Observations: $o_1 = (p_1 < 2), o_2 = (p < 2)$ and $o_3 = (p3 < 2), o_4 = (p4 < 2), o_5 = (age < 70)$
- Clinical criterion: $cc_1 = (o_1, o_2), cc_2 = (o_3), cc_3 = (o_4, o_5)$
- Clinical criteria: $CC(typical - AD) = \{cc_1\}$, not $\{cc_2, cc_3\}$

This provides some kind of freedom in the description of the clinical criteria, for example in the description of the phenotypes or more complex diagnosis. For example:

- $CC(AD) = \{CC(typical - AD), CC(atypical - AD)\}$,
- $CC(MCI) = not\{CC(x), x \in Dementia\}$,$cc(impairment < 3, impairment > 0)$.

Eventually, such definitions can be revised and updated for the technology that will deal with the inclusion of the biomarkers. Additionally there is the power of the dealing with a case in a personalized manner, as the case can be described through such definitions, and possible interventive measures can be defined through counterfactuals.

Complexity Sources of noise in the process can increase the complexity. Among the common problems are:

- test noise: the tests do not necessarily reflect the level of impairment, as they might be affected by several factors (ex. was the subject in a good mood, is possible retesting?),
- diagnosis noise: the diagnosis might not be stable (ex. transitory state, mixed diagnosis),
- missingness: some of the information might be missing (ex. are the biomarkers available for AD?).

These allow for the clinician's input and verdict to be based on a more holistic perspective. To be able to handle these, a relatively wide assumption of noise needs to be taken, allowing for errors or milder levels of judgement. For example, while the research criteria suggests that memory is a later problem for DLB, having it early might be related to the late observation because of compensatory mechanisms. While realistic,

this kind of process is more difficult to be represented and validated. These cases can be handled through counterfactuals post-analysis, and the description for that is provided later in this chapter.

Predictors Based on the protocols and the literature, the following are the features that are involved in the diagnosis of Dementia. `Dx` stands for diagnostic value in relation to the other Dementia disorders group, while `Ph` stands for a predictor involved in the phenotypic differentiation.

Table 2: Category of predictors relevant for the disorders.

Type	Predictor category	Value	Notes
MCI			
ci	health-history	Ph	defines etiology
impairment	impairment, progression	Dx	general predictor of MCI
npa	memory	Ph	for a-MCI
npa	executive-f., language, visuospatial	Ph	for na-MCI
nimg(vol)	CSF, temporal, cingulate, precuneus	Dx&Ph	
nimg(pat)	limbic	Dx	
AD			
ci	demographics, health-history	Ph	age, gender
impairment	impairment	Dx	vs. MCI
impairment	onset	Dx	vs. FTLD vs VD
nps	psychotic	Dx	vs. DLB - hallucinations
npa	memory	Dx&Ph	general predictor of AD
npa	language, executive-f.	Dx	vs. FTLD in presence of early age
npa	attention, executive-f., visuospatial	Dx	vs. DLB
nimg(vol)	hippocampus, entorhinal cortex	Dx&Ph	
nimg(pat)	CSF	Dx	amyloid or tau deposition
nimg(pat)	temporal, parietal	Dx	hypometabolism
DwMD			
physical	motor	Dx&Ph	motor dysfunctions profiling is important
physical	somatic	Dx&Ph	sleep disturbances - REM sleep - DLB or MSA
physical	sensory	Ph	sensory loss - CBD
nps	psychotic	Ph	hallucinations - DLB
nps	others	Ph	pseudobulbar - PSP

Type	Predictor category	Value	Notes
nps	personality	Dx	co-occurrence with bv-FTD, PNFA
npa	executive-f., visuospatial, psychomotor	Dx	
npa	attention	Ph	DLB
npa	memory	Ph	DLB and PSP defining
nimg(vol)	limbic, brain-stem	Dx&Ph	DLB atrophy
nimg(vol)	basal-ganglia, pons, cerebellum	Dx&Ph	MSA atrophy
nimg(pat)	basal-ganglia	Dx	tau presence
FTLD			
ci	demographics	Ph	gender - male SD and bvFTD, female PNFA
impairment	onset	Dx	earlier than other progressive dementias
npa	personality	Dx	general predictor of FTLD,
npa	executive-f.	Ph	behavioral, set shifting, flexibility, abstract reasoning
npa	language	Ph	semantic (anomic) vs PNFA (non-fluent) vs bv
npa	visuo-spatial, memory	Dx	vs. AD, spared
nimg(vol)	frontal, temporal, cingulate, insular	Dx&Ph	vs. AD, mainly bv-FTD
nimg(vol)	motor, insula, operculum	Ph	PNFA
nimg(pat)	frontal, temporal	Dx	hypometabolism
VaD			
ToDO			

2.5. Computational models

The modeling of Dementia can be done through building a model that uses clincial features, as the ones mentioned in the Clinical criteria, or Clinical examination sections. The models are simply trying to understand the relation between the predictors (features) and the result. We provide the model with some initial data to train it, and present later some new data, and see whether the model can accurately predict the resulting value. A simple linear model is shown below at (I), where x_{i_j} is the input, θ_i is the parameter, and y_i is the resulting wanted prediction. The training is done to find the best parameters through optimizing a loss function, and reducing the parameter space and avoiding overfitting through regularization, as shown in (II). The combination of the loss (Λ) and regularization (Ω) is called the objective function (obj). The Mean Squared Error (MSE) is an example of a loss function used for regression, and the Lasso regularization, as shown in (III) and (IV).

$$\text{(I)}\ \ \hat{y}_i = \frac{1}{N}\sum_j \theta_i x_{i_j}$$

$$\text{(II)}\ \ obj(\theta) = \Lambda(\theta) + \Omega(\theta)$$

$$\text{(III)}\ \ \Lambda(\theta) = \sum_{i=1}(\hat{y}_i - y_i)^2$$

$$\text{(IV)}\ \ \Omega(\theta) = \lambda \sum_{i=1} |\theta_i|$$

The algorithms are various, starting from simple methods like Logistic Regression, to more advanced ones like Ensemble models or Neural Networks. For each of the methods there are several variations of the models that might make them more compatible for specific tasks.

Ensemble tree models The tree models are based on trees that separate the decision space based on features (see figure below). For an introduction to trees check sklearn documentation (Pedregosa et al. 2011).

Table 3: The different types of machine learning models.

Model Type	Variations	Value for Diagnostic models
Logistic Regression	Regularized with Lasso, Ridge	Reduce the feature space
Support Vector Machines	SVC-based, SVM	Classification standard
Distance-based	KNN, K-means	Useful for imputation
Tree models	Decision Tree, Random Forest	Intuitive results
Ensemble models	XGBoost, LightGBM, AdaBoost	Good results with intuitive bases
Bayesian networks	Naive Bayes, Bayesian belief networks	Probabilistic integration
Neural networks	CNN-based, RNN-based, Graph-NN	Advanced black-box modelling

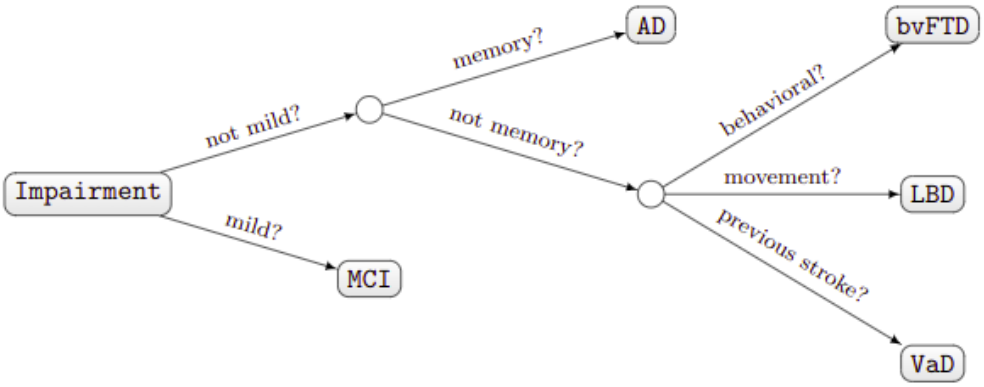

Figure 2: An example decision tree on Dementia.

A variation of trees are the classification and regression trees (CART) (Chen and Guestrin 2016). CART keeps a score for each separation and decision that allows the combination of many of them into an ensemble model. The ensemble based scoring is shown in (V), and the objective function described in (VI).

$$\text{(V)}\quad \hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F}$$

$$\text{(VI)}\quad \text{obj}(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \omega(f_k)$$

Models of Dementia For example, the tree models are more explainable, while the Neural Networks might perform better. The resulting power of these models is dependent on the type of data that is being processed too. Recent literature reviews and meta-analyses show the extensive work done on models that try to diagnose Dementia, and their preference for different types of datasets.

- The neuropsychological measurements alone have a differentiating power for the HC-MCI-AD progression, but not a prognostic power (Duke Han et al. 2017; Battista et al. 2020).
- Multimodal prediction of Dementia through deep learning and medical imaging makes possible early detection of the HC-MCI-AD spectrum diseases (Ahmed et al. 2019).
- Personalization of the diagnostic process can be based on the open datasets, and inclusion of various biomedical information is valuable (Kumar et al. 2021).

Additionally, the latest models have achieved positive results that can have predictive power. These models can be integrated to geriatrics departments for improving the patient care through correct early diagnostic evaluation or prognostic evaluation. For the diagnosis several ensemble models have proven useful (Bogdanovic, Eftimov, and Simjanoska 2022), and for prognostic evaluation the temporal recurrent neural networks trained on longitudinal data have shown the first results (Jung, Jun, and Suk 2021). After these models have proven

useful for an extensive range of tests and datasets, they can be integrated to the clinical practice for specific correctable decisions.

Private models Models built from clinical data can leak the training data, which makes the task of learning from closed datasets (for example the extensive datasets of governmental institutions) risky. A mathematical solution to this problem is the Differential Privacy, which in combination to Explainable Boosting Machines (EBMs), can provide a long-term solution to the reproducibility and extensibility of the pipelines of analysis.

2.6. Explaining models

The explainability of machine learning in the case of diagnosis of Dementia because it allows insight in the decision-making process of the models. One valuable contribution of such models is the ability to find predictor interactions that might have not been observed before as important to the diagnostic process. In this sense, exploring the interpretability of the model can be insightful to the clinical decision making process. There are several ways to explain how the models learn, among them SHapley Additive Explanation (SHAP), Locally Interpretable Model-agnostic Explanations (LIME), Partial Dependence plots, and more are being developed.

SHAP is a method based on game theory that considers each predictor as a player of several games, and makes evaluations on the impact of each such predictor on the game's result (Lundberg and Lee 2017). The algorithm tries to do a fair allocation of importance to each player. The main usefulness of the SHAP method is its interpretabiity of the model both in the global and local sense.

2.6.1. Example

For the sake of explanation, let's take the case where four diseases have little overlapping, and the patients show a defined set of symptoms. A clinical diagnostic process would take note of these predictors: `Age`, `Impairment`, `Memory`, and `Gait`. The protocol for diagnosis of diseases for such minimalistic process based on the existing guidelines would be as follows:

- Disease 1: `Memory` problems is the main predictor, shown mostly with increased `Age`.
- Disease 2: `Gait` problems are the main predictor, no `Memory` problems.
- Disease 3: young `Age` is the main predictor, and shows high level of `Impairment`.
- Disease 4: a lower degree of `Impairment` than the other diseases is the main predictor.

A model would find these relationships and would correctly predict the diseases if a new subject is provided. The simple relationship `Disease 1` ~ `Memory` is easier to measure than the `Disease 4` ~ `Impairment` level. Linear or logistic regression is able to capture the first relationship, while DecisionTrees would be more able

to represent the conditional relationship in the second one. What if we have a patient that at a young `Age` shows a low level of `Impairment`? Then we need a more advanced model, but in the moment that the models become more complicated, the importance of each predictor, and the combined effect that some predictors might have, is more difficult to measure.

2.6.2. Logic of SHAP

In this context, SHAP values are helpful. They take all the predictors and measure their impact on the diagnostic result in terms of coalitions. So, the predictors are players of these games, and they contribute at changing (or maintaining) the default result of the game. The default result of the toy-example game is the most prevalent disease, for example `Disease-4`. While it might be easy to measure the impact of a single predictor like `Impairment` on the diagnostic result, it is more difficult to measure how the combination of `Age`, `Impairment` and `Memory` affect the diagnosis in collaboration, besides their impact on their own.

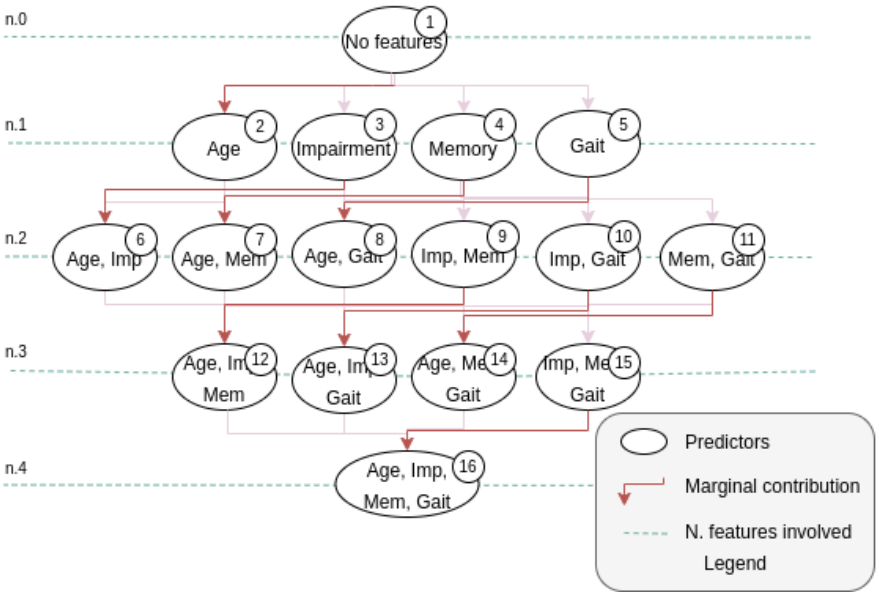

Figure 3: Combination of predictors for building SHAP.

For each of the combinations that are provided in the figure, SHAP calculates a predictive model that starts from same data and parameters and considers only the predictors in the node. Each node in the graph in the figure shows a SHAP internal model built with the mentioned predictors. Each edge shows the marginal contribution (MC) of each predictor.

For example, the effect of `Age` is measured from the difference in impact between groups of predictors not containing `Age`, and then adding `Age` to that group. As shown in the figure, starting from no predictors and adding `Age` is one (`MC(Age){1,2}`), but also starting from `Imp`, `Gait` and adding `Age` also reflects the marginal

contribution of `Age` (`MC(Age){10,12}`). Adding all these contributions allows for an overall overview of how does adding `Age` affect the groups of predictors.

To have a balanced distribution of the impact, the impact of each model is weighted through the coefficient `w` that represents the relative contribution for the level. For example, each predictor at level 0 to level 1 has the weight 1/4, as they are in total 4 edges. In the second level there are two edges for each combination of predictors, so 6*2=12. For a full example check the example (**???**). Eventually the function for measuring SHAP for `Age` has the following form, that can generalize for each predictor, and for each coalition of predictors.

```
SHAP(Age) = 1/4 x MC(Age){1,2}
  + 1/12 x MC(Age){3,6} + 1/12 x MC(Age){4,7} + 1/12 x MC(Age){5,8}
  + 1/12 x MC(Age){9,12} + 1/12 x MC(Age){10,13} + 1/12 x MC(Age){11,14}
  + 1/4 x MC(Age){15,16}
```

Plots Visualizations of the impact of the predictors can be observed through the different plots that aim at the global and local explainability. Plots aiming at explaining the example relation are found in the end of this chapter.

- Global explainability plots: The SHAP value provides knowledge about how much a certain predictor impacts the prediction of the model. The visualization tools used tend to be intuitive and help build trust to the machine learning models decision-making process (Lundberg et al. 2020).
- Local explainability plots: These plots reflect some kind of decision making similar to a clinician. The graph is an effect (x) by feature graph (y). The effect of each feature is added strating from the bottom, with the features in the top being the most impactful ones.

2.6.3. Counterfactual explanations

Understanding the decisional process of the model is important but does not provide actionable intelligence. For example, if a subject is in the state of MCI and a model can predict that their diagnosis will change soon into AD, the question that comes up is: how can we prolong the process? For that we can use the counterfactual explanations (CE). CE are examples of the smallest actions to be taken so that the prediction can change (Wachter, Mittelstadt, and Russell 2017). These examples are the closest as possible to the provided case, and the possible changes can be provided. The counterfactual generation is based on an optimization algorithm that minimizes a loss function based on several requirements. An example method is introduced by Dandl and colleagues (Dandl et al. 2020), where they define the requirements for plausible counterfactuals:

- the prediction should be close to our desirable prediction,
- the counterfactual should be as close to the initial case,
- the changes should include a small amount of features,
- the feature changes should be likely (based on the dataset).

These methods can be valuable for the clinician to take action to improve a certain area for the patient or to train another area for the patient to not deteriorate. In the case of MCI mentioned above, training the cognitive side might help into maintaining the current diagnosis, while some actions like increasing the cardio-vascular medications can deteriorate the condition. In combination with SHAP these explanations can be used for understanding what the best strategy for an institution (for example a hospital) is to apply trainings at a subgroup of patients.

2.7. Hypotheses

The aims of the thesis are to handle issues like cost (test time and money), validity (protocols), and potential interventions (counterfactuals). The process creates a parallel process to the diagnostic process and analyses the combination of factors through explainable machine learning to define the clinical protocols mathematically, for making them testable and changeable. This kind of expansion of the method should allow an iterative improvement, and expansion of the models, and some explainability to the expert knowledge.

With the data from the open datasets, like the NACC dataset, we can have a chance at looking at the statistical models. They can add some insights on the disorders, at least be comparable to the clinical criteria.

Hypothesis 1: Computational models based on extensive clinical examination (NACC dataset) of patients of the HC-MCI-AD spectrum are able to diagnose similarly to the ones built from a restricted clinical assessment and neuroimaging (ADNI dataset).

Hypothesis 2: Predictors selected from automatic selection of the model are similar to the ones defined in the clinical protocol for the diagnosis of diseases: MCI, AD, DLB, FTD, or VaD (NACC dataset).

Hypothesis 3: The counterfactual examples of mis-diagnosed cases from the computational models creates subgroups for possible interventions for the most prevalent diseases.

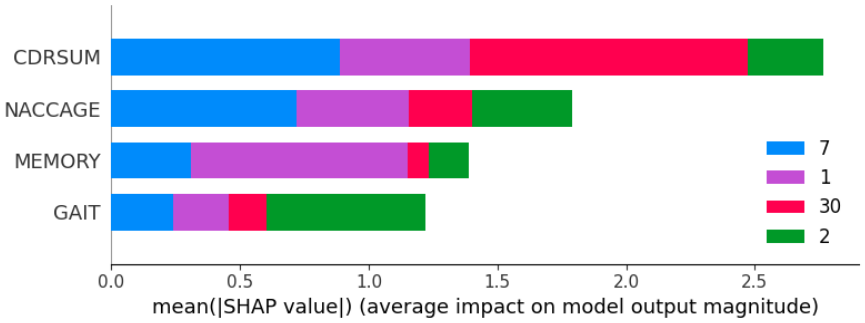

Figure 4: Feature Importance plot. The most important features plot can be an initial interaction with the existing plot.

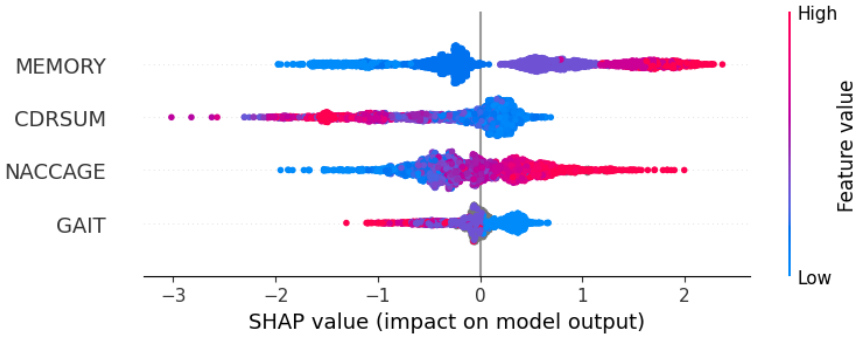

Figure 5: Summary plot. In the summary plot, this SHAP value of every feature inserted in the prediction is set in relation with the value of the respective feature. It provides an overview about which feature is important (SHAP value on x-axis) if it has a certain value (y-axis).

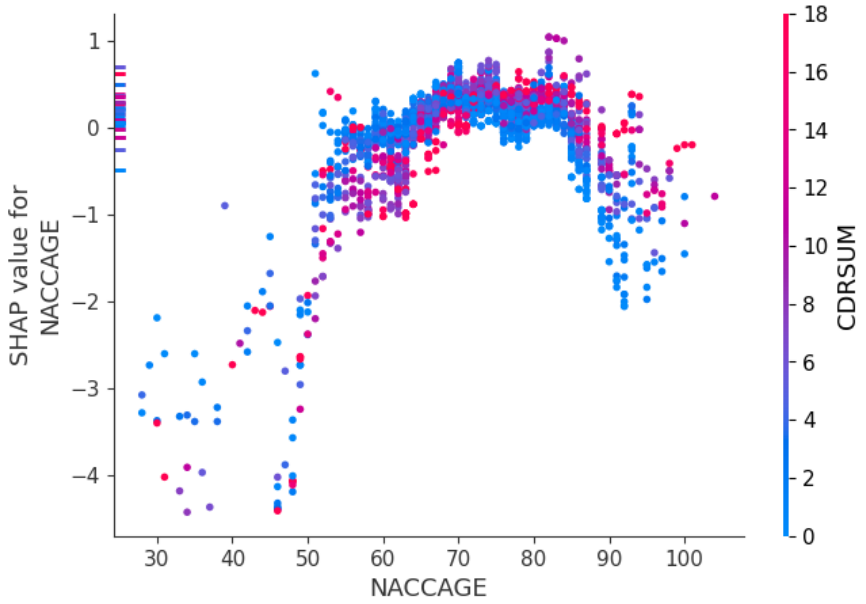

Figure 6: Dependence plot. In a next step, one single of the features is picked out and plotted with its value on the y-axis and its SHAP value on the x-axis. The dependence plot zooms into one single feature and its behavior in the prediction depending on its value.

Force plot. For a single observation, it shows the accumulated importance of the positive and negative features in a horizontal bar. Additionally it can be done for a group of instances, and this can be useful for visualizing mistaken observations from the model.

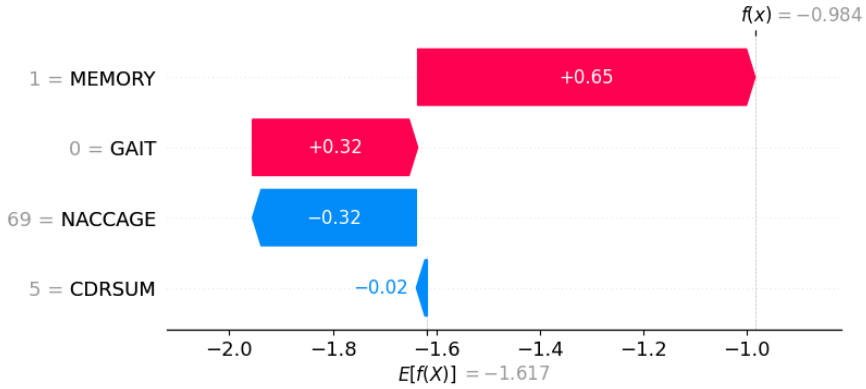

Figure 7: Waterfall plot. How did the model's output deviate from its expected output? In a cascade-like shape the features contributing positively and negatively to the model's output value are visualized above each other ordered from lowest importance at the bottom to highest importance at the top.

Decision plot. The more intuitive explanations of the models can be done using the decision plots, and The features are sorted by importance of the subsample shown in the graph, and possibly can be hierarchically clustered.

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

Table 4: 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

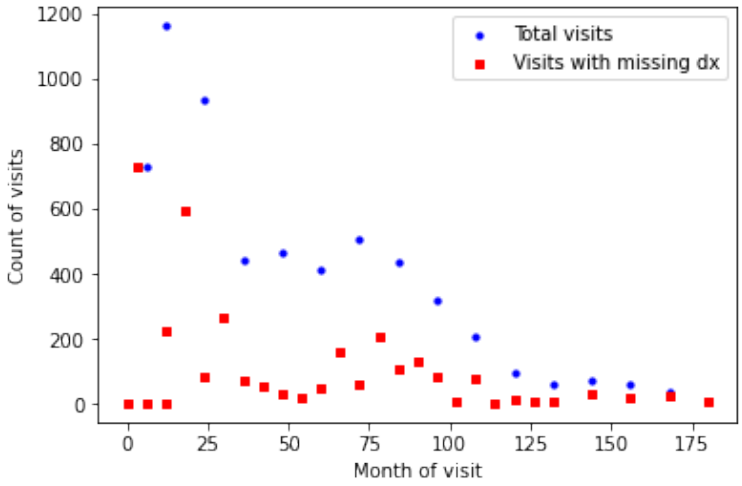

Figure 8: Missing diagnoses in ADNI.

- 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.

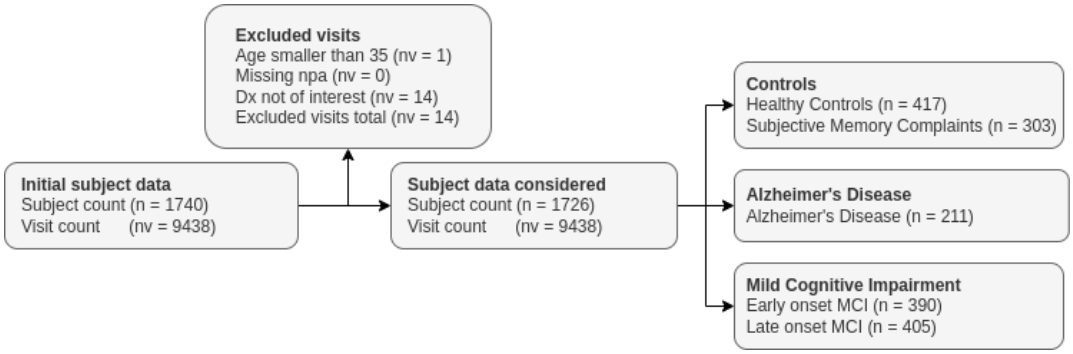

Figure 9: ADNI Participant Flow

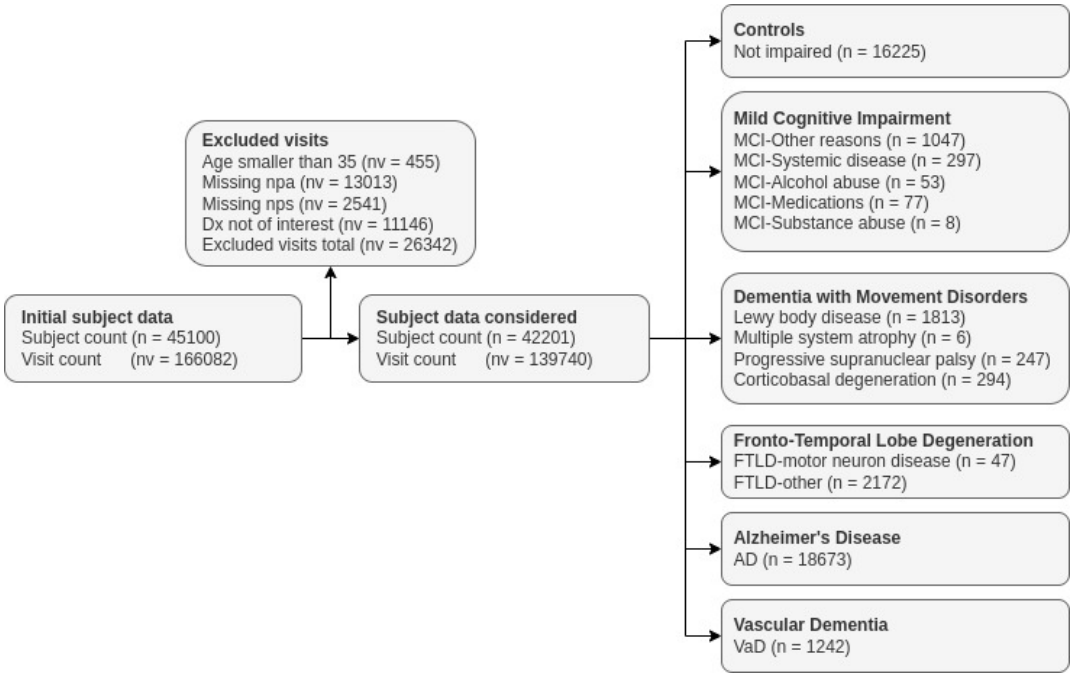

Figure 10: NACC Participant Flow

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 5: 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]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.

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

  - 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 7: 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 11: 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 8: 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.

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.

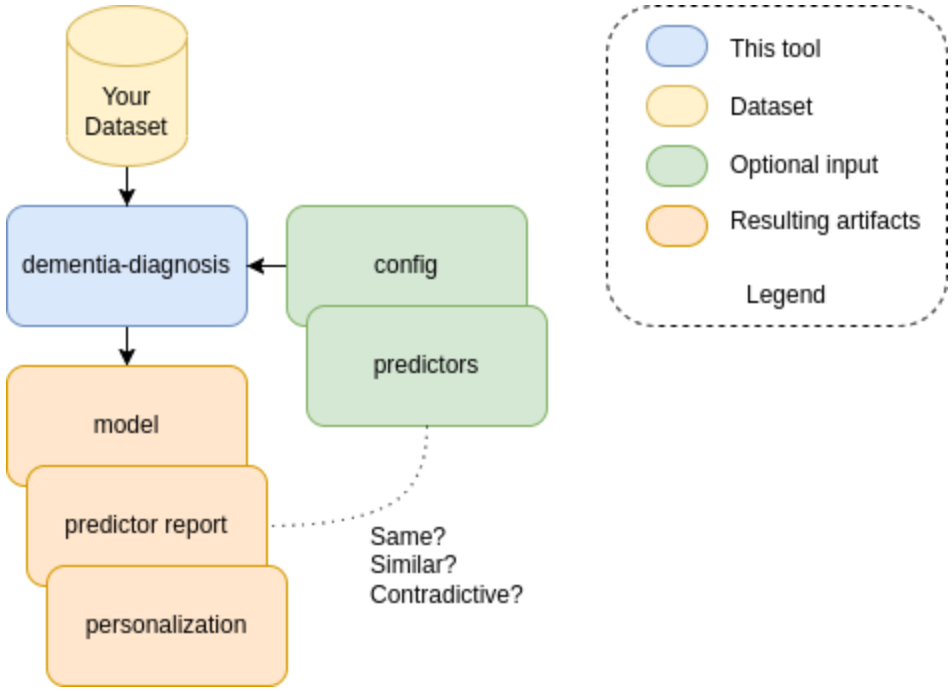

Figure 12: Processing tool.

The `config.json` and the `variables.csv` files are added in the supplementary material. The code for reproducing this work can be found online.[5] 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.

[5] 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 9: Dataset descriptions.[6]

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

[6]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`.

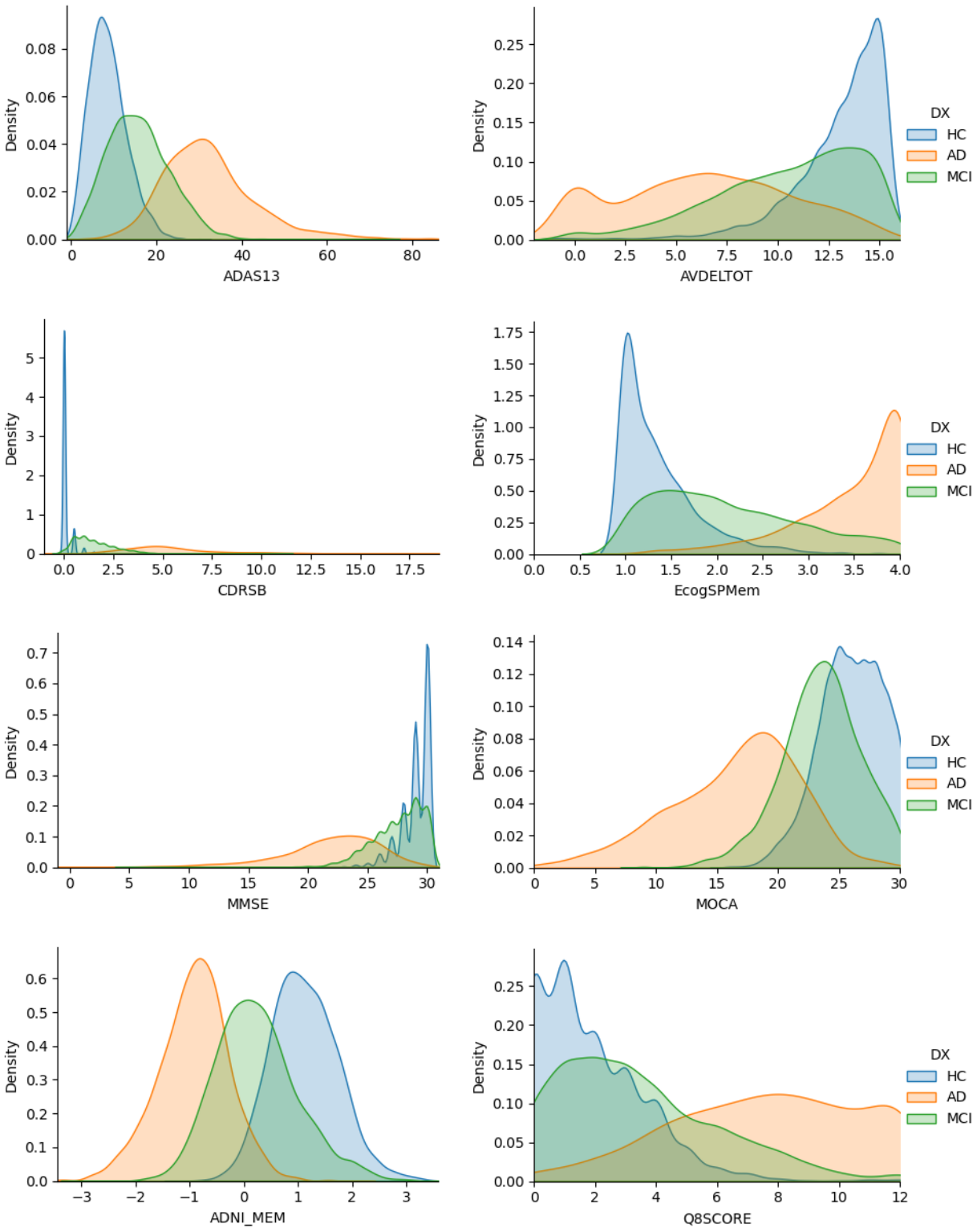

Figure 13: Cognitive scores related to ADNI.

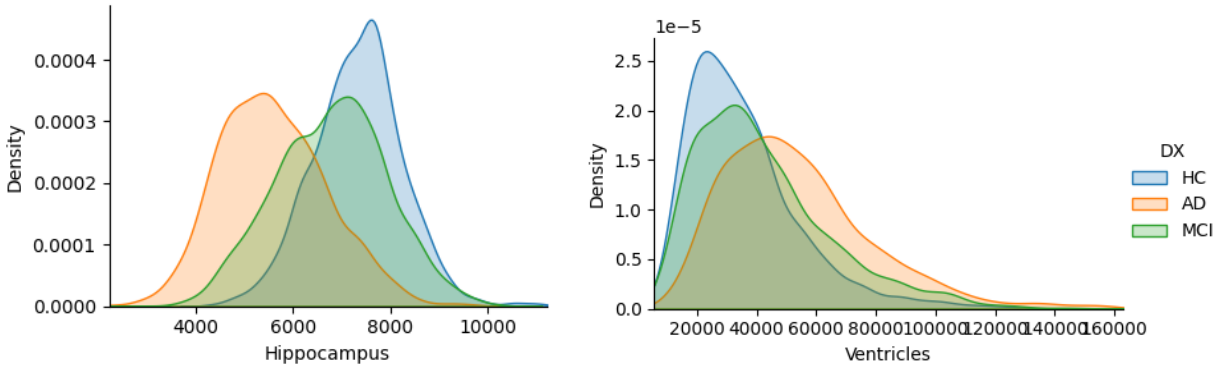

Figure 14: Hippocampus and Ventricles volumes for ADNI.

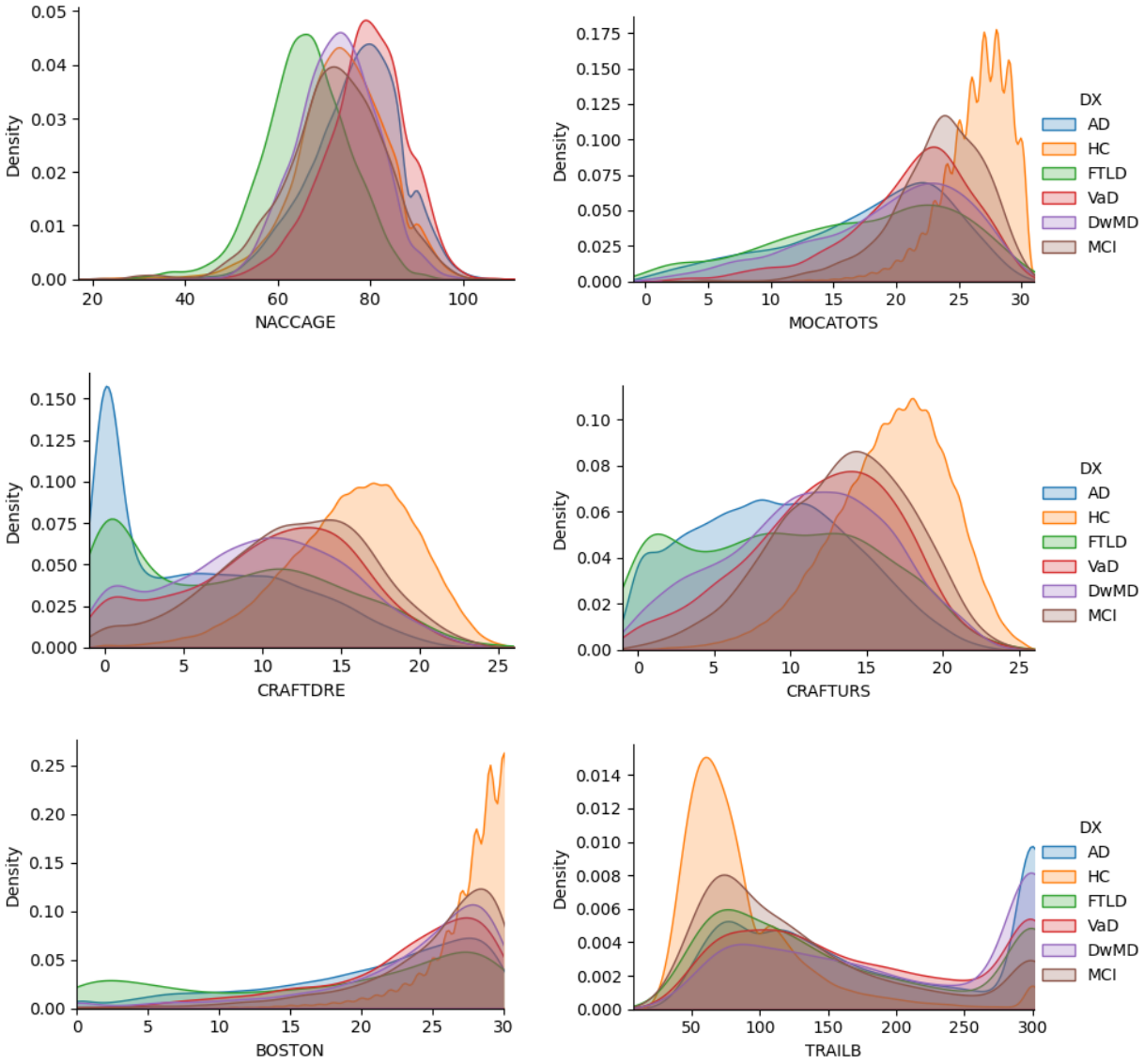

Figure 15: NACC numerical data samples.

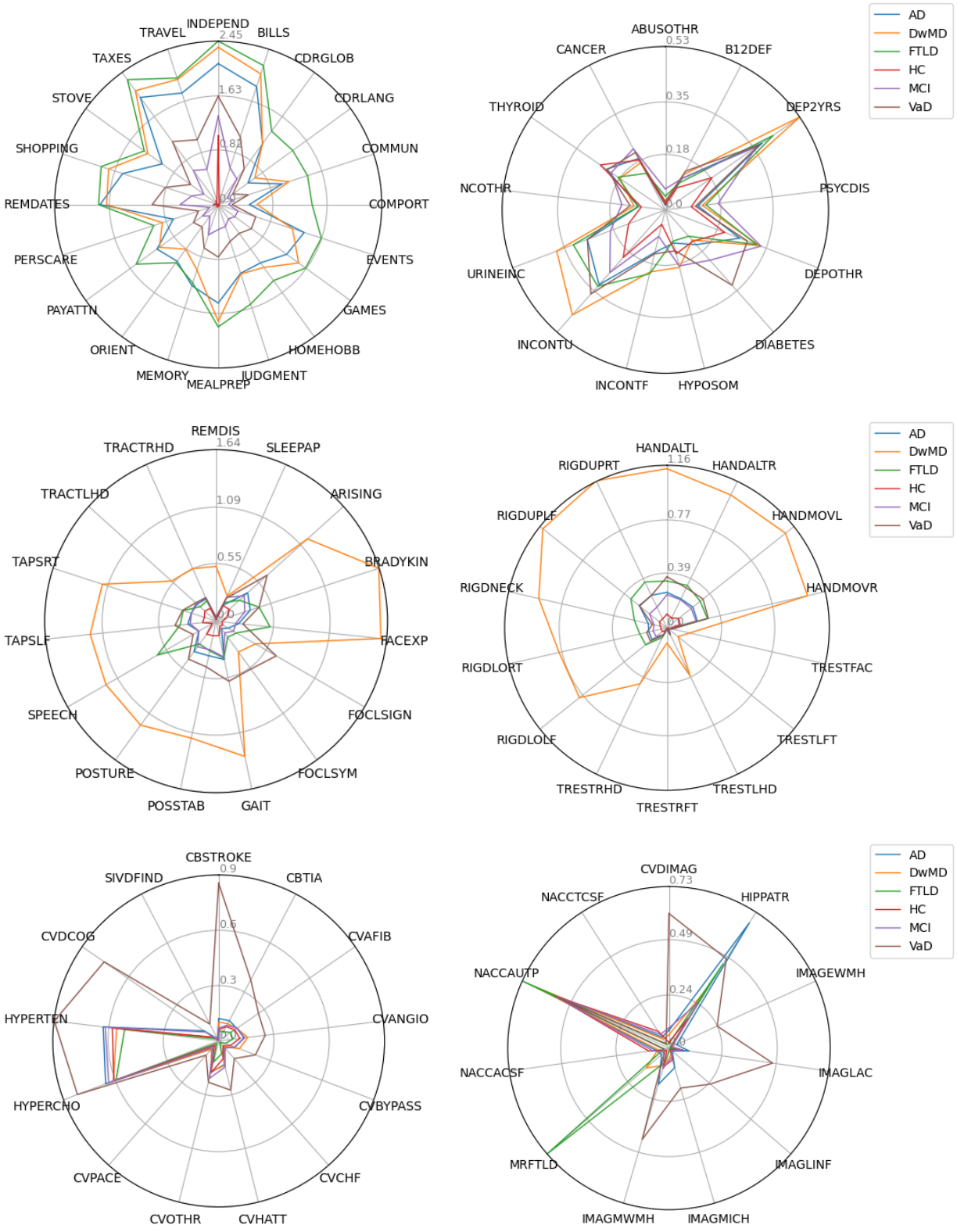

Figure 16: Impairment, medical history, parkinsonism symptoms, tremor-related symptoms, cardio-vascular history, and neuroimaging categorical data in the NACC dataset in that order.

Table 10: 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 11: 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

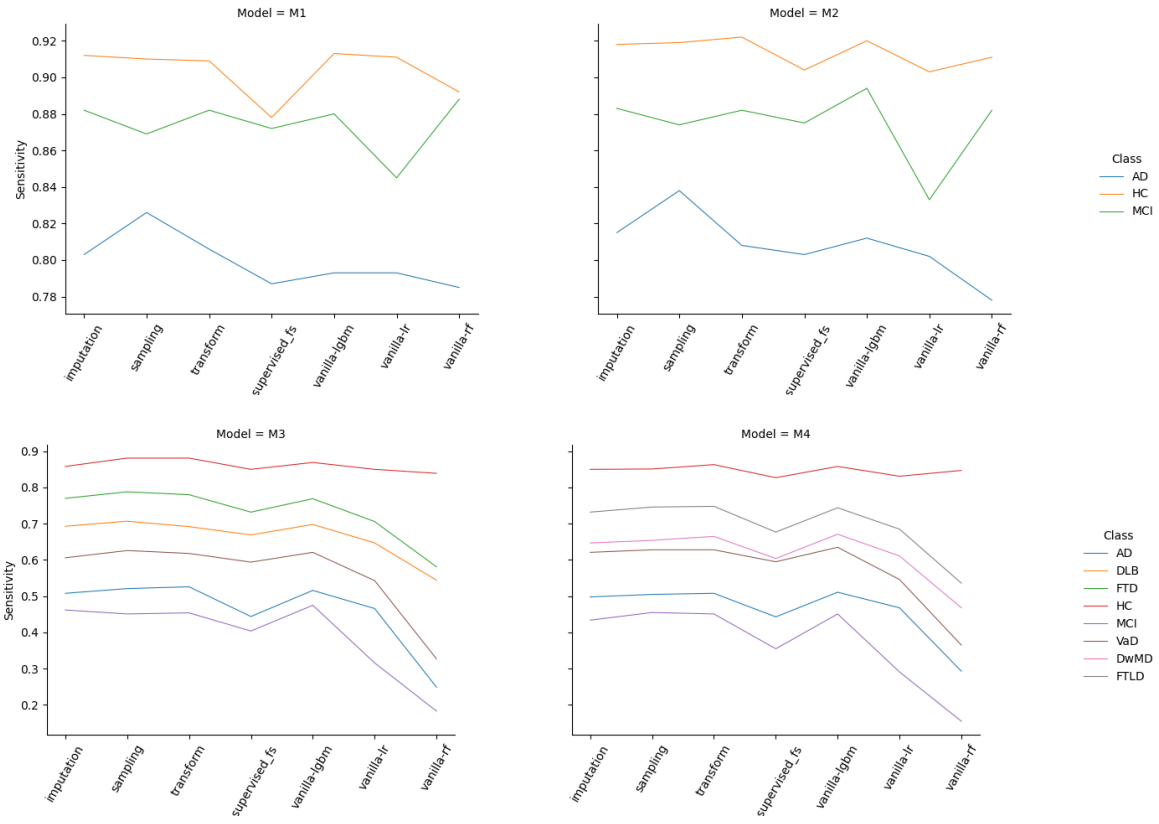

Figure 17: ADNI and NACC 'Sensitivity' scores for each class. As it can be observed, the better models are the LGBM models, as they remain sensitive for MCI in ADNI models and for MCI and DwMD for NACC.

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.

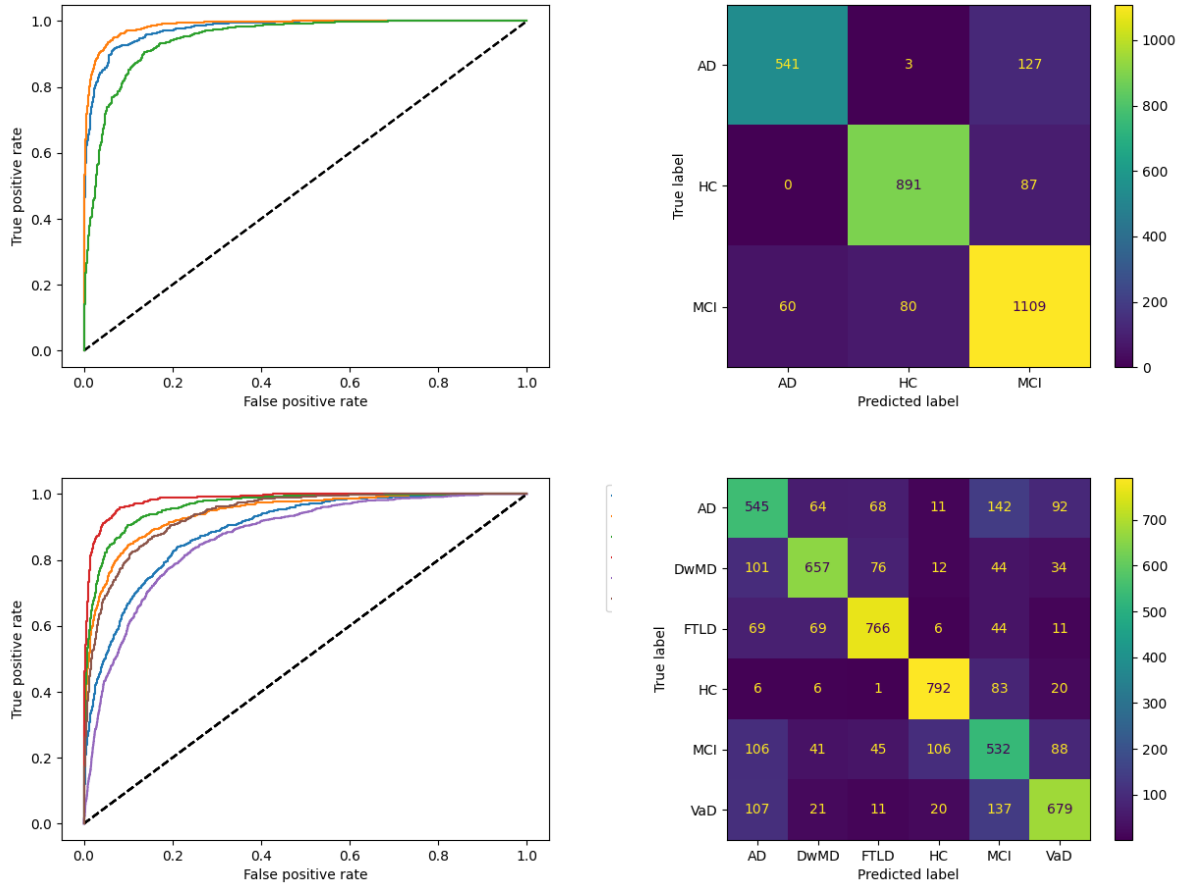

Figure 18: ROC curves and the confusion matrix for the M2-ADNI (first row) and M4-NACC (second row) vanilla-lgbm models. The ROC curves combine the true positives versus false positives, while the confusion matrix shows how the models do in terms of combined results. In the first row ADNI graphs, with 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.

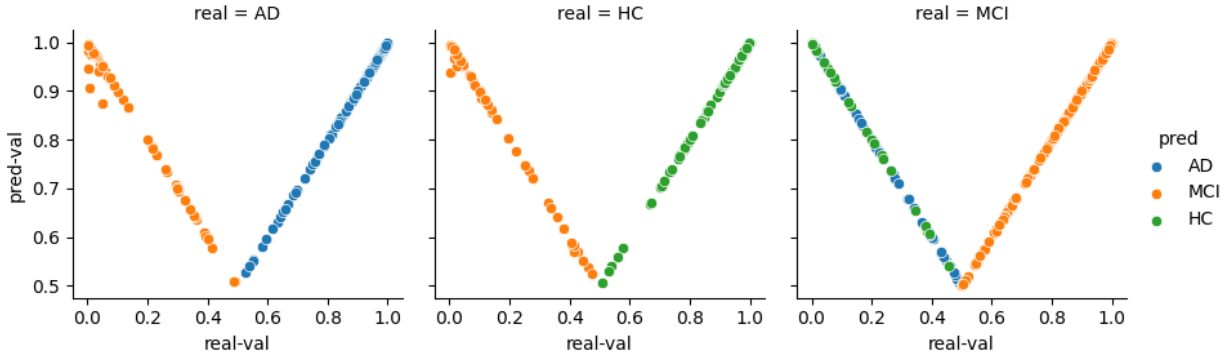

Figure 19: M2 distribution of mis-classifications.

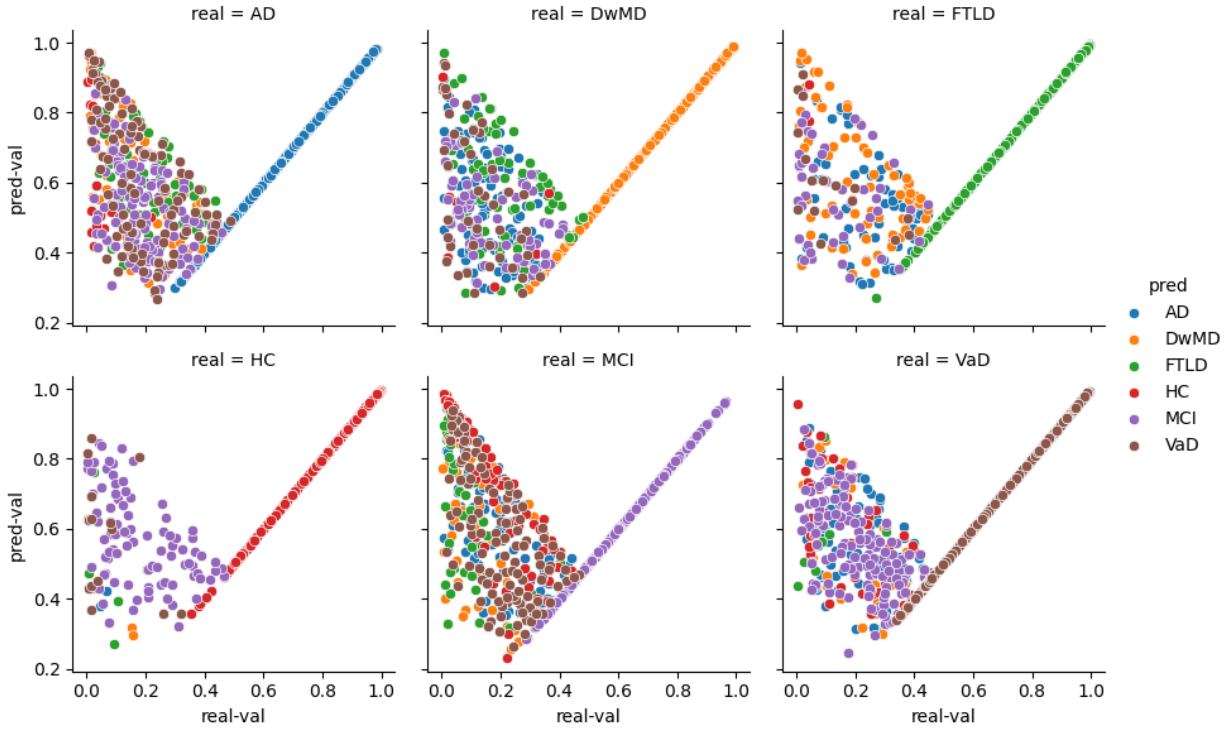

Figure 20: M4 distribution of mis-classifications.

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.

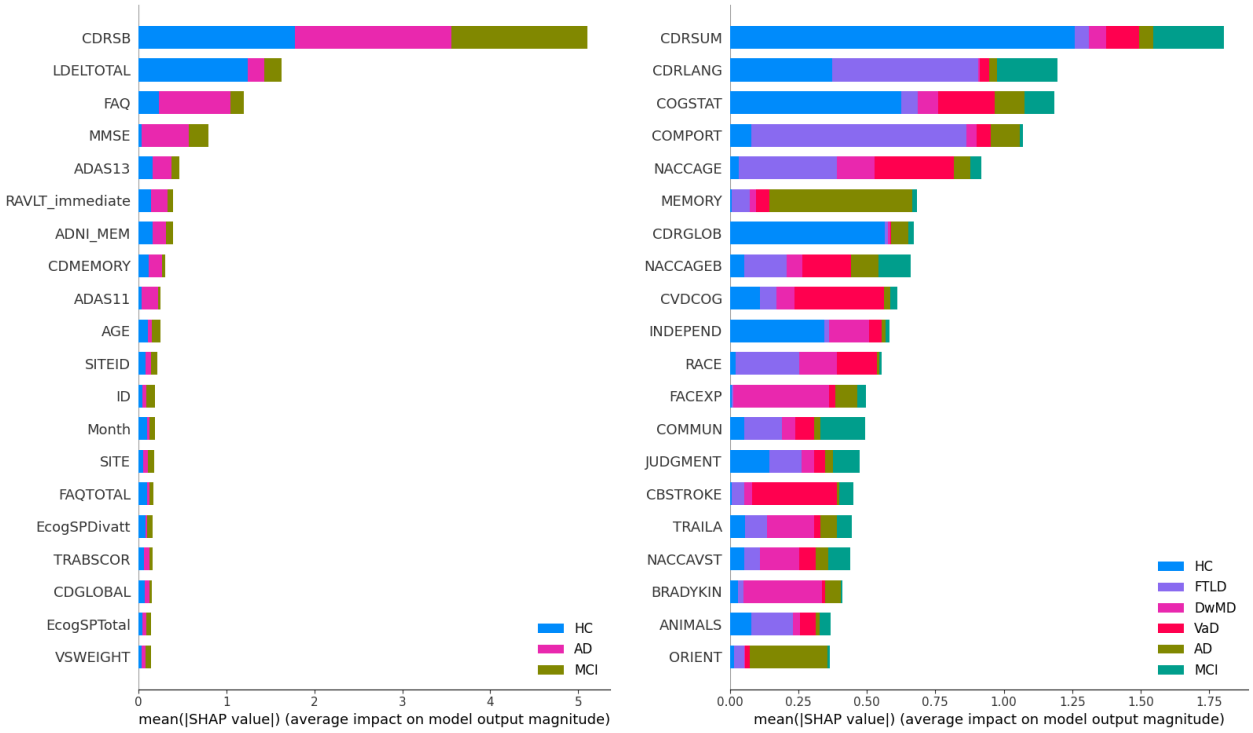

Figure 21: Feature importance for ADNI and NACC models respectively, ordered by impact.

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.[7] 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.

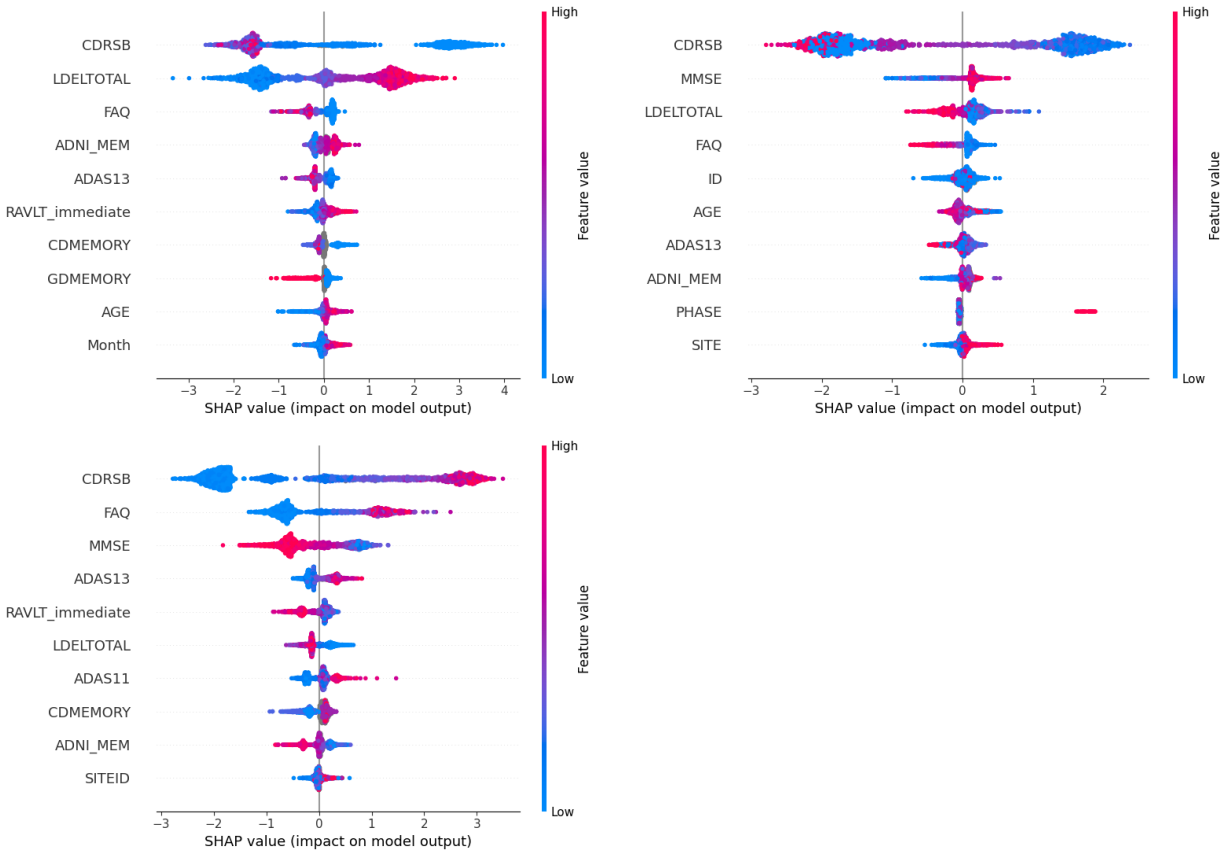

Figure 22: ADNI summary plot for HC, MCI, and AD respectively.

[7] 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.[8] 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.

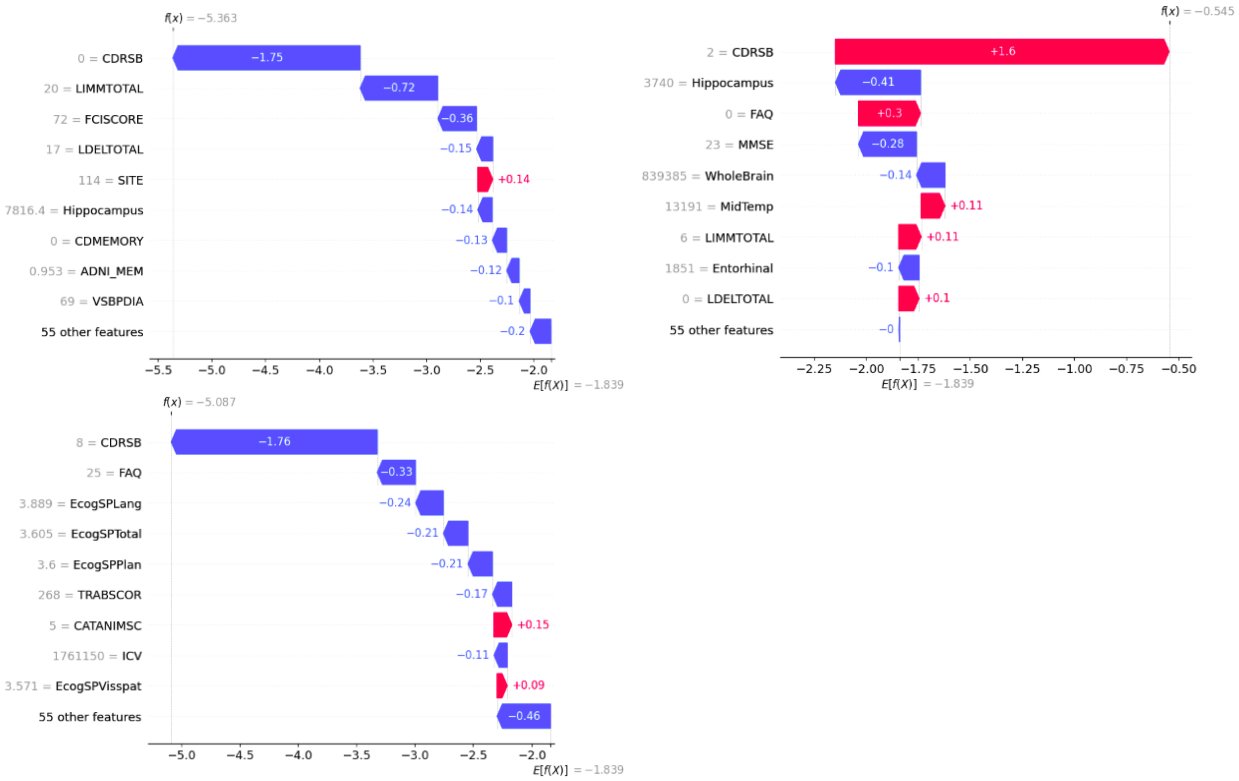

Figure 23: ADNI single case model decision descriptions for HC, MCI, and AD.

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`.

[8] 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.

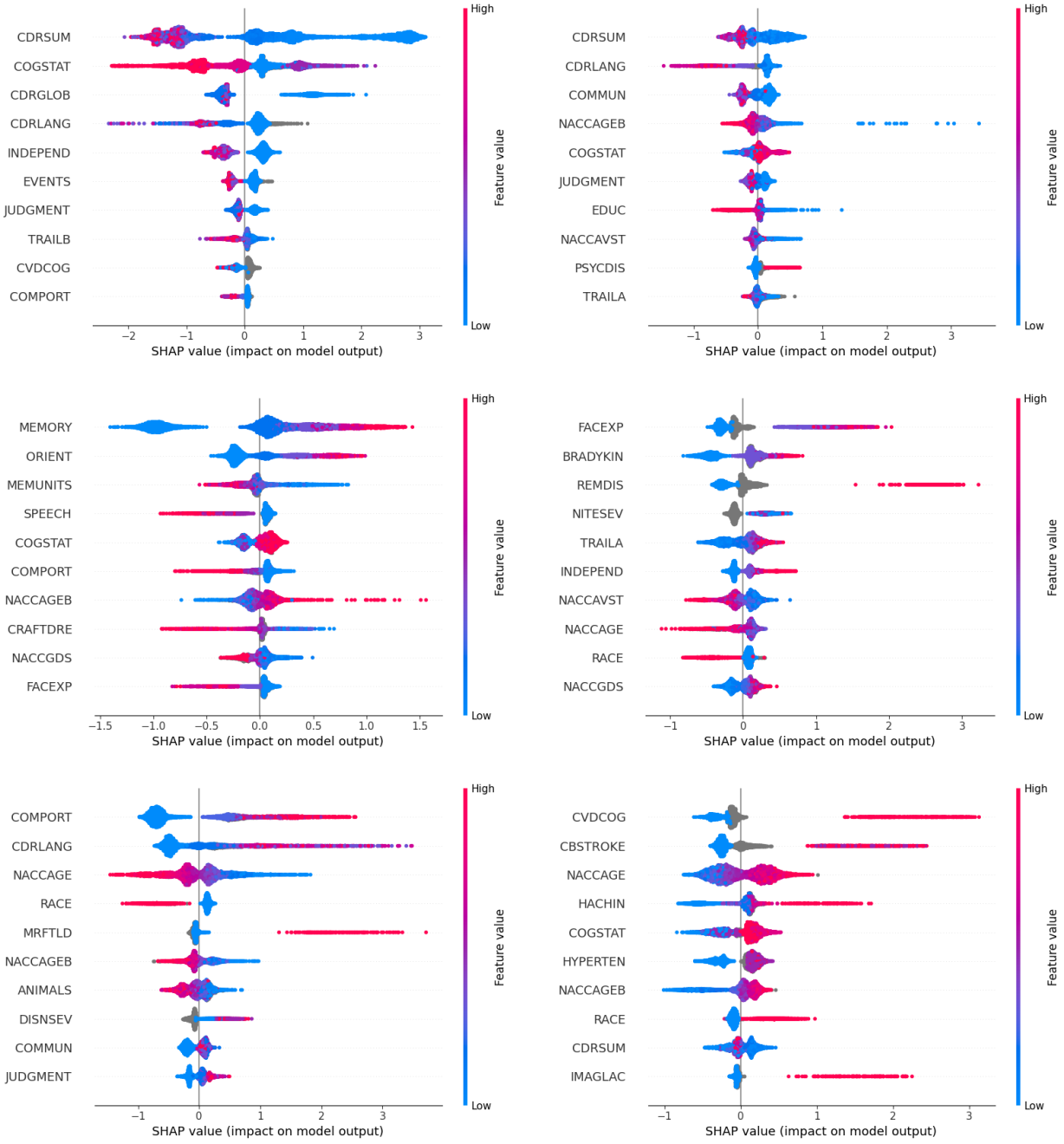

Figure 24: NACC summary plot for HC, MCI, AD, DwMD, FTLD, and VaD in that order.

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`.

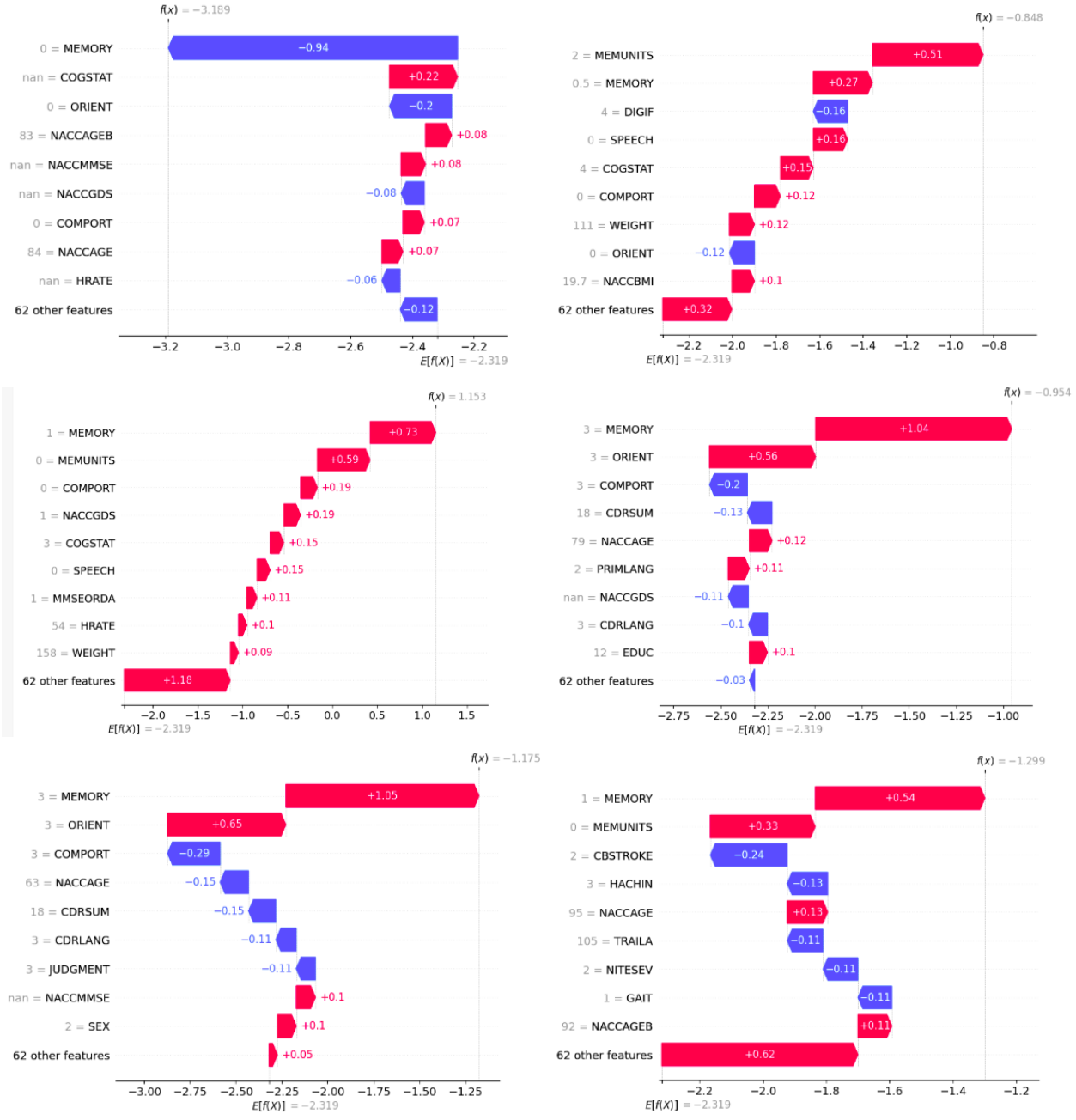

Figure 25: NACC single case model decision descriptions for HC, MCI, AD, DwMD, FTLD, and VaD.

Chapter 5. Discussion

This works attempts to revise the clinical criteria for Dementia using explainable models of machine learning. It additionally establishes a baseline process against which the models can be benchmarked and tested. The process was tested using two open datasets: ADNI and NACC, and four models were built. The Alzheimer Disease Neuroimaging Initiative dataset was smaller, and it included subjects of the {HC-MCI-AD} spectrum. The National Alzheimer's Coordinating Center dataset was similarly extensive in terms of data types, and it included a wider range of Dementia cases. This work did not intend to achieve the highest performance, but to evaluate whether the models can be useful in to re-defining the boundaries of existing protocols. The following are lessons learned:

The model building is highly reproducible, and the presence of a pipeline allows for further expansion. These models can also be developed for other datasets. Light Gradient Boosting Method tends to do better than the other models (such as the RanodmForestModel or LogisticRegression), making it more reliable in the diagnostic process. Additionally, the LGBM pipeline requires almost no preprocessing, which makes it easier for the process of model-building to avoid introducing artifacts. The results for A DNI are as good as the literature describes.

When building models using detailed demographic and neuropsychological data performs as good as having a more restricted dataset including neuroimaging too. The features that are important for the definition of the classes {HC, MCI, AD} are related mostly to the neuropsychological tests like CDRSB, FAQ, MMSE, and do not require the neuroimaging data.

The clinical criteria defined in the protocols can be reproduced using these models, in both cases. For the diseases that are in NA C C the predictors like {CDRSUM, MEMORY, COGSTAT} are defining for the {HC, MCI, AD} spectrum, {CDRLANG, COMPORT} are relevant for F T L D, {FACEXP, REGDUPRT, REMDIS} are relevant for DwMD, and {CBSTROKE, HACHIN, CVDCOG} are relevant features for VaD. Additionally, for the case of MCI which is more complicated than the other disorders the models learn the diagnosis by exclusion.

This work was exploratory, delineating the boundaries of the diseases, and see whether machine learning can provide some insight. The clinical criteria for the different types of Dementia are based on the separation between core criteria and supportive criteria. To date it is standard to have a structure for the diagnostic process that includes conditional logic. In all diseases an essential component is a progressive cognitive decline (I. F. McKeith, Boeve, and Dickson 2017; Dubois et al. 2014; Winblad et al. 2004). Through the insights

provided by these models there may be multiple feedback loops in the disease improvement of the protocols of these complex diseases. This process is illustrated below.

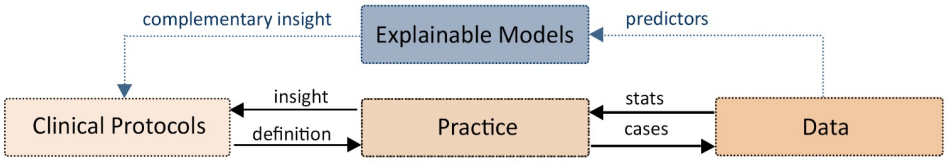

Figure 26: Protocol update cycles

The importance of some neuropsychological tests in the diagnostic process can also be seen explicitly. While seeing how the distributions of the test scores related to the several diseases can be useful (as in Section 4.1.), understanding how the tests interact with each other is also important. This can allow for the interaction of the conditional format described in the protocols. For example, the interaction and conditionals described in the work of (I. F. McKeith, Boeve, and Dickson 2017) needs to allow for such updates.

Limitations This work has several limitations. First, the datasets have longitudinal data, but they are used as cross-sectional data in the models. The core feature for the definition of Dementia is cognitive decline. This work does not make use of the information that lies in the progression of the subjects, and treats the cases as single shot learning. Although the models created have prognostic value, they do not aim to be predictive in that sense. Nevertheless, the problems that may arise from not considering the relation between the visits of same patients were considered and used to improve the validity of the models. This can be seen in the grouping of the validation method.

Also, during the cleaning of the datasets, there has been a unification of the missigness by reducing the codes provided from the datasets into NaN. These codes sometime include useful information, still a further processing step would be out of scope for this work. As described in chapter 3, missing clinical data can be distinguished in the NAC C and ADNI datasets. In the NACC dataset does provide some information on the type of missingness, and such information could be used to create better and more detailed strategies. Overall, the models tend to not show big differences, and the resulting model does well with no imputation needed.

What was lacking of the validation of the pipeline is an extensive validation of parameters. While in this thesis there has been a concentration on the methodology, the extensive testing of the parameters has not been done, mainly due to computational reasons and the sizes of the datasets. This suggests that the models

that have been predicted are probably in the local maxima, so the predictions can be improved. This technical limit was known beforehand.

Besides these, there is a significant accuracy drop between the models built from ADNI and NACC, mainly related to the fact that NACC models are doing a six class classification, while the ADNI models are doing a 3 class classification. Even though this difference exists, it is of importance to understand why the line is blurry, and is the process dificult, or could the model do much better? While in general the boundaries between the diseasess are clearer through explainability analysis, more must be done in increasing the predictive power of the model.

Recommendations This work provides an in-depth look of the insights that can be given from the models. Further development of the method would require a fix for the limitations, and an extensions to the depth of the models used here. Also, there is a further interest in seeing how these models can be integrated to the daily practice of dealing of subjects with Dementia. While the datasets used are extensive, the models allow for an easier way to standardize the process by selecting the more significant tests. This standardization can help the professionals and caregivers to use the time of more eficiently and have more reliable and reproducible results.

Additionally, this process of standardization and lowering of the number of tests makes the handling of Dementia diagnosis more manageable. In the context of the two datasets, the tests defined in their clinical protocols are defined in detail in their study definition. The variation of tests used, their specificity, the equivalence scores, and the additional information can be provided (in an electronical format) for the tests. This kind of information transfer can allow reproducibility for the clinical application. Such open-data initiatives are not possible right now because of the licenses that the tests have, but should be allowed in the future.

Besides the diagnostification, there might be need to apply similar pipelines for the prognostic models. Among the latest methods for dealing with longitudinal data is the use of deep learning models that can account for the relation between the several visits by the same subject. Such models can make such inferences, and then make projections about the short-term risk of progression. This work did not attempt at dealing with such models as it was out of scope. When this work will be extended for clinical usage, the deep learning models can be tested too.

Chapter 6. Conclusions

The main objective of the work was to test whether there can be alternative ways to define the importance of the clinical criteria for Dementia using machine learning. Two main hypotheses were established based on the existing data and the practical necessity of the models. A pipeline was built allowing the extensive testing of several computational models, and the better pipeline was picked (LGBM) which combined minimal preprocessing and good sensitivity results. Then a further development of such work to include the descriptive statistics for each Dementia disease, and related single case analysis was presented. The comprehensive analysis suggests that explainable machine learning models can be useful at providing complementary insights to the existing protocols.

Supplementary Information Access to the ADNI data can be requested at https://adni.loni.usc.edu/data-samples/access-data/. Access to the NACC data can be requested at v. The NACC dataset was created for the purpose of this study The code for this thesis can be found at https://github.com/DorenCalliku/thesis.

Software and hardware The pipeline, data processing and reporting are performed in a Linux machine, with four memory cores, Intel i5 processor, 16GB Ram, and python 3.8.2 installed. A full list of the requirements of the python running environment can be found the Appendix C. The analyses were all run in a single thread for measuring computational cost. The codes are available on request.

Funding No funding was received for this work.

References

Ahmed, Md Rishad, Yuan Zhang, Zhiquan Feng, Benny Lo, Omer T. Inan, and Hongen Liao. 2019. "Neuroimaging and Machine Learning for Dementia Diagnosis: Recent Advancements and Future Prospects." *IEEE Reviews in Biomedical Engineering* 12: 19–33. https://doi.org/10.1109/RBME.2018.2886237.

Albert, M S, S T DeKosky, and D Dickson. 2011. "The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease." *Alzheimers Dement* May;7(3):2. https://doi.org/10.1016/j.jalz.2011.03.008.

Alzheimer Europe. 2019. "Dementia in Europe yearbook 2019: Estimating the prevalence of dementia in Europe." https://www.alzheimer-europe.org/sites/default/files/alzheimer_europe_dementia_in_europe_yearbook_2019.pdf.

Armstrong, M J, I Litvan, A E Lang, T H Bak, K P Bhatia, B Borroni, A L Boxer, et al. 2013. "Criteria for the diagnosis of corticobasal degeneration." *Neurology* 2335. https://doi.org/10.1212/WNL.0b013e31827f0fd1.

Battista, Petronilla, Christian Salvatore, Manuela Berlingeri, Antonio Cerasa, and Isabella Castiglioni. 2020. "Artificial Intelligence and Neuropsychological Measures: The Case of Alzheimer's Disease." *Neuroscience and Biobehavioral Reviews* 114 (July): 211–28. https://doi.org/10.1016/j.neubiorev.2020.04.026.

Bogdanovic, Bojan, Tome Eftimov, and Monika Simjanoska. 2022. "In-depth insights into Alzheimer's disease by using explainable machine learning approach." *Sci. Rep.* 12 (1).

Brain. 2009. *Riluzole treatment, survival and diagnostic criteria in Parkinson plus disorders: the NNIPPS study.* Edited by G Bensimon, A Ludolph, Y Agid, M Vidailhet, C Payan, and P N Leigh. *Brain.* NNIPPS Study Group. https://doi.org/10.1093/brain/awn291.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." https://doi.org/10.48550/ARXIV.1603.02754.

Dandl, Susanne, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. "Multi-Objective Counterfactual Explanations." In *Parallel Problem Solving from Nature PPSN XVI*, 448–69. Springer International Publishing. https://doi.org/10.1007/978-3-030-58112-1_31.

Duke Han, S., Caroline P. Nguyen, Nikki H. Stricker, and Daniel A. Nation. 2017. "Detectable Neuropsychological Differences in Early Preclinical Alzheimer's Disease: A Meta-Analysis." *Neuropsychology Review* 27 (4): 305–25. https://doi.org/10.1007/s11065-017-9345-5.

Gilman, S, G K Wenning, and P A Low. 2008. "Second consensus statement on the diagnosis of multiple system atrophy." *Neurology* 26;71(9):6. https://doi.org/10.1212/01.wnl.0000324625.00404.15.

Hernandez, Monica, Ubaldo Ramon-Julvez, and Francisco Ferraz. 2022. *Explainable AI toward understanding the performance of the top three TADPOLE Challenge methods in the forecast of Alzheimer's disease diagnosis*. Vol. 17. 5 May. https://doi.org/10.1371/journal.pone.0264695.

Josephs, Keith A. 2007. "Frontotemporal Lobar Degeneration." *Neurologic Clinics* 25 (3): 683–96. https://doi.org/10.1016/j.ncl.2007.03.005.

Jung, Wonsik, Eunji Jun, and Heung-Il Suk. 2021. "Deep Recurrent Model for Individualized Prediction of Alzheimer's Disease Progression." *NeuroImage* 237 (August): 118143. https://doi.org/10.1016/j.neuroimage.2021.118143.

Kouri, Naomi, Yari Carlomagno, Matthew Baker, Amanda M Liesinger, Richard J Caselli, Zbigniew K Wszolek, Leonard Petrucelli, et al. 2014. "Novel mutation in MAPT exon 13 (p.N410H) causes corticobasal degeneration." *Acta Neuropathol.* 127 (2): 271–82.

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.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. "From Local Explanations to Global Understanding with Explainable Ai for Trees." *Nature Machine Intelligence* 2 (1): 2522–5839.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

McKeith, I F, B F Boeve, and D W Dickson. 2017. "Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium." *Neurology* 89: 88–100.

McKeith, I G, T J Ferman, and A J Thomas. 2020. "Research criteria for the diagnosis of prodromal dementia with Lewy bodies." *Neurology* 28;94(17): https://doi.org/10.1212/WNL.0000000000009323.

McKhann, Guy M, and David S Knopman. 2011. "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic

guidelines for Alzheimer's disease." *The Diagnosis of Dementia Due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease* 7 (3): 263–69. https://doi.org/10.1016/j.jalz.2011.03.005.

Nichols, Emma, Cassandra E. I. Szoeke, Stein Emil Vollset, Nooshin Abbasi, Foad Abd-Allah, Jemal Abdela, Miloud Taki Eddine Aichour, et al. 2019. "Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016." *The Lancet Neurology* 18 (1): 88–106. https://doi.org/10.1016/S1474-4422(18)30403-4.

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.

Organisation, World Health. 2017. "Global action plan on the public health response to dementia 2017 - 2025." http://www.who.int/mental_health/neurology/dementia/action_plan_2017_2025/en/.

Park, Denise C., and Patricia Reuter-Lorenz. 2009. "The Adaptive Brain: Aging and Neurocognitive Scaffolding." *Annual Review of Psychology* 60 (1): 173–96. https://doi.org/10.1146/annurev.psych.59.103006.093656.

Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Persson, Jonas, Lars Nyberg, Johanna Lind, Anne Larsson, Lars-Göran Nilsson, Martin Ingvar, and Randy L. Buckner. 2006. "StructureFunction Correlates of Cognitive Decline in Aging." *Cerebral Cortex* 16 (7): 907–15. https://doi.org/10.1093/cercor/bhj036.

Petersen, R C. 2004. "Mild cognitive impairment as a diagnostic entity." *J Intern Med* 256: 193–4.

Pini, Lorenzo, Michela Pievani, Martina Bocchetta, Daniele Altomare, Paolo Bosco, Enrica Cavedo, Samantha Galluzzi, Moira Marizzoni, and Giovanni B. Frisoni. 2016. "Brain atrophy in Alzheimer's Disease and aging." *Ageing Research Reviews* 30: 25–48. https://doi.org/10.1016/j.arr.2016.01.002.

Rascovsky, Katya, John R. Hodges, David Knopman, Mario F. Mendez, Joel H. Kramer, John Neuhaus, John C. Van Swieten, et al. 2011. "Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia." *Brain* 134 (9): 2456–77. https://doi.org/10.1093/brain/awr179.

Salmon, David P, and Mark W Bondi. 2009. "Neuropsychological assessment of dementia." *Annu. Rev. Psychol.* 60 (1): 257–82.

Salmon, D P, R G Thomas, M M Pay, A Booth, C R Hofstetter, L J Thal, and R Katzman. 2002. "Alzheimer's disease can be accurately diagnosed in very mildly impaired individuals." *Neurology* 59 (7): 1022–8.

Salthouse, Timothy A. 1996. "The Processing-Speed Theory of Adult Age Differences in Cognition." *Psychological Review* 103 (3): 403–28. https://doi.org/10.1037/0033-295x.103.3.403.

Stern, Yaakov. 2009. "Cognitive reserve." *Neuropsychologia* 47 (10): 2015–28. https://doi.org/10.1016/j.neuropsychologia.2009.03.004.

Tosi, Giorgia, Carolina Borsani, Stefania Castiglioni, Roberta Daini, Massimo Franceschi, and Daniele Romano. 2020. "Complexity in neuropsychological assessments of cognitive impairment: A network analysis approach." *Cortex* 124: 85–96. https://doi.org/10.1016/j.cortex.2019.11.004.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3063289.

Wardlaw, J M, E E Smith, and G J Biessels. 2013. "STandards for ReportIng Vascular changes on nEuroimaging (STRIVE v1). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration." *Lancet Neurol* Aug;12(8): https://doi.org/10.1016/S1474-4422(13)70124-8.

Winblad, B, K Palmer, M Kivipelto, V Jelic, L Fratiglioni, L.-O. Wahlund, and R C Petersen. 2004. "Mild cognitive impairment - beyond controversies, towards a consensus." *Report of the International Working Group on Mild Cognitive Impairment. J Intern Med* 56: 240–6.