SLaCAD: A Spoken Language Corpus for Early Alzheimer's Disease Detection

Shahla Farzana,[1] **Edoardo Stoppa,**[1] **Alex Leow,**[1] **Tamar Gollan,**[2] **Raeanne Moore,**[2] **David Salmon,**[2] **Douglas Galasko,**[2] **Erin Sundermann**[2] and **Natalie Parde**[1]
[1]University of Illinois Chicago
[2]University of California San Diego
[1]{sfarza3, estopp2, weihliao, parde}@uic.edu
[2]{tgollan, r6moore, dsalmon, dgalasko, esundermann}@health.ucsd.edu

Abstract

Identifying early markers of Alzheimer's disease (AD) trajectory enables intervention in early disease stages when our currently-available interventions are most likely to be beneficial. Research has shown that alterations in speech, as well as linguistic and semantic deviations in spontaneous conversation detected using natural language processing, manifest early in AD prior to some other observed cognitive deficits. Recent studies show that cerebrospinal fluid (CSF) levels serve as useful early biomarkers for identifying early AD, but CSF biomarkers are challenging to collect. A simpler alternative that has seen very rapid development is based on the use of plasma biomarkers as a blood draw is minimally invasive. Associating verbal and nonverbal characteristics from speech data with CSF and plasma biomarkers may open the door to less invasive, more efficient methods for early AD detection. We present SLaCAD, a new dataset to facilitate this process. We describe our data collection procedures, analyze the resulting corpus, and present preliminary findings that relate measures extracted from the audio and transcribed text to clinical diagnoses, CSF levels, and plasma biomarkers. Our findings demonstrate the feasibility of this and indicate that the collected data can be used to improve assessments of early AD.

Keywords: CSF biomarker, plasma biomarker, early diagnosis, Alzheimer's disease

1. Introduction

Alzheimer's disease (AD) is an irreversible neurodegenerative disease with a long preclinical phase where pathology in the brain develops gradually over many years before diagnostic symptoms manifest (Blennow et al., 2006; Jack Jr. et al., 2018). In the preclinical AD phase, there is evidence of AD pathogenesis via biomarkers but no clinical symptoms. Pharmaceutical and non-pharmaceutical approaches such as lifestyle changes promoting brain health might be helpful to slow or delay cognitive decline in cases detected early (Kivipelto et al., 2017), but detection of the earliest, preclinical AD phase is challenging (Håkansson et al., 2018).

In the prodromal stage of AD called mild cognitive impairment (MCI), mild cognitive deficits, typically in the domain of learning and memory, manifest clinically; however, everyday function is not impaired. Recent clinical trials of disease-modifying agents suggest that therapeutics work best if started at early stages (Sharma, 2019), making it a high priority to develop diagnostic tools for early AD stages that are sensitive and less invasive, costly and time-burdensome than those currently available (e.g., brain imaging or biomarker assays in cerebrospinal fluid (CSF) or blood).

Biomarkers of the hallmark AD pathologies, amyloid-β (Aβ) plaques and neurofibrillary tangles comprised of phosphorylated tau, can be measured in the CSF and plasma and can be detected years before the onset of clinical symptoms. The most commonly measured AD biomarkers are $A\beta_{42}$, $A\beta_{42}/A\beta_{40}$ ratio, total-tau (tTau), the phosphorylated tau protein at epitope 181 ($pTau_{181}$) and the ratio of either tTau or $pTau_{181}$ to $A\beta_{42}$. High concentrations of $pTau_{181}$, tTau, and the tTau (or $pTau_{181}$)/$A\beta_{42}$ ratio levels and low levels of $A\beta_{42}$ and the $A\beta_{42}/A\beta_{40}$ ratio reflect greater pathological burden in the brain. These biomarkers have consistently predicted subsequent progression to AD in cognitively unimpaired and MCI participants (Rostamzadeh et al., 2022; Breno S. O. Diniz and Forlenza, 2008; Ferreira et al., 2014). Although recent advances in the use of plasma AD biomarkers have reduced barriers to collection, the need for phlebotomy and assay cost still restricts their widespread use. Additionally, many individuals with AD pathology in the brain not develop Alzheimer's disease (Driscoll and Troncoso, 2011; Arenaza-Urquijo and Vemuri, 2018), indicating that other markers of the earliest cognitive change in the AD trajectory are needed. Clinical markers of this coupled with biomarkers would make a powerful combination in detecting individuals likely on the AD trajectory.

Changes in spontaneous spoken or written discourse have been observed early in the course of AD, and possibly prior to MCI (Forbes-McKay and Venneri, 2005a; Garrard et al., 2004a). There is ev-

LREC-COLING 2024, pages 14877–14897
20-25 May, 2024.

idence that different connected speech tasks may be sensitive to different linguistic features in early AD diagnostics (Clarke et al., 2021). Recent studies have suggested that computational analysis of spontaneous speech could be a rapid, low-cost, scalable, and non-invasive screening tool for early AD. We introduce the **S**poken **L**anguage **C**orpus for **A**lzheimer's **D**isease detection (SLaCAD), a new dataset to advance research in this area. We describe our comprehensive approach to elicit spoken discourse from 91 older, mostly cognitively normal, participants. Participants completed language tasks, resulting in 7.5 hours of recorded and transcribed speech data across participants. The transcripts of the language recordings and the extracted linguistic and acoustic features along with clinical diagnosis (cognitively normal, MCI, AD dementia), comprehensive cognitive testing, and AD biomarkers (for the subset of participants indicated in this paper) will be available to the scientific community through data requests made to the National Alzheimer's Coordinating Center. The availability of early AD biomarkers to characterize preclinical AD in this mostly cognitively normal sample paired with spoken discourse will facilitate research towards early AD screening. Our contributions are:

- We present SLaCAD, collected from 91 participants through autobiographical interviews in clinical laboratory settings.
- We derived language transcripts from recordings and correspondingly extracted linguistic and acoustic features from the transcripts. These data will be available to other researchers by request.
- Using SLaCAD, we relate these features with CSF and plasma AD biomarkers.

Through these analyses, we identify linguistic and acoustic features that correlate with AD biomarkers in a mostly cognitively normal sample. This is promising because it represents an innovative way to potentially detect subtle signs of AD pathology and risk before cognitive impairment becomes apparent. We detail our data collection procedures, analyses, and findings in the remainder of this paper.

2. Background

2.1. Spoken Language Corpora for Detecting Early AD

Several publicly available or requestable spoken language datasets relevant to early AD exist, each with different sample characteristics and data availability. DementiaBank (Becker et al., 1994) contains audio recordings of neuropsychological tests administered to healthy participants and patients with diagnosed dementia. It includes 300 language samples from 188 participants with cognitive decline and 242 samples from 99 cognitively normal, older adults. Out of the 300 interviews from participants with cognitive decline, 43 interviews were classified as from participants with MCI and 257 as from participants with possible/probable AD. However, DementiaBank does not include any CSF or plasma biomarkers.

The Framingham Heart Study (Wawrzyniak, 2020, FHS) has language recordings/data and diagnostic labels that are available upon request. Audio was recorded during a picture description (PD) Task. Apart from the diagnostic labels, FHS has plasma amyloid-β (Aβ_{42}) (Romero et al., 2020) and plasma total-Tau (tTau) biomarkers (Pase et al., 2019). A Swedish corpus (Jonell et al., 2021) also contains multimodal data (gaze, speech, and facial gestures) from 25 participants, as well as diagnostic labels (AD, MCI or control), the CSF Aβ_{42} and phosphorylated tau (p-tau) biomarkers, and the Montreal Cognitive Assessment (MoCA) Memory Index Score (MoCA-MIS) for each participant. Findings from studies on this data demonstrated correlations between speech biomarkers and AD biomarkers.

Recent research (Verfaillie et al., 2019) found that 63 individuals with subjective cognitive decline (SCD) from a memory clinic and high amyloid burden uttered fewer specific words during an English-language spontaneous speech task. Another English-language study (Mueller et al., 2021), using cookie theft picture description data from the Wisconsin Registry for Alzheimer's Prevention[1] with 255 participants (57 amyloid positive and 198 amyloid negative), showed that participants with positive amyloid status demonstrated poor performance over time in linguistic parameters (i.e., low vocabulary richness) compared to participants with negative amyloid status in a cohort of cognitively healthy individuals. Finally, other recent research (Hajjar et al., 2023), using privately collected data from 92 cognitively unimpaired (40 Aβ positive) and 114 impaired (63 Aβ positive) participants, found that lexical-semantic features extracted from spoken English picture descriptions were significant in the detection of positive Aβ status using machine learning techniques.

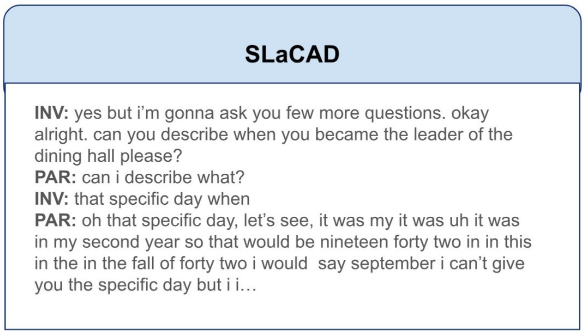
2.2. Speech and Language Markers for Early AD Detection

Evidence suggests that changes in spoken or written language can occur early in AD, possibly before MCI (Forbes-McKay and Venneri, 2005b; Garrard et al., 2004b; Ahmed et al., 2013). These lan-

[1] https://wrap.wisc.edu/

guage abilities are controlled by brain regions like the parieto-temporal and temporal lobes, which are often affected early in AD. In practical terms, this can result in difficulties finding words, slower speech, hesitancy, and trouble understanding language. Many studies have used NLP to extract linguistic and semantic features for detecting AD progression (Slegers et al., 2018; Mueller and Turkstra, 2018; Voleti et al., 2020). Some studies have also compared the sensitivities of different speech sampling approaches (e.g., picture description or semi-structured interviews) to early AD detection (Seyed Ahmad Sajjadi and Nestor, 2012), finding that discourse samples elicited from semi-structured interviews contain more fillers (e.g., "uh" and "um"), incomplete utterances, and grammatical function words than picture description tasks. In contrast, picture descriptions allowed for the capture of more semantic errors, such as substituting the word "dog" for "cat" (Seyed Ahmad Sajjadi and Nestor, 2012). Given these findings, it seems likely that the task used to elicit spoken discourse not only affects the accuracy of the classifier but also the nature of the distinguishing features.

Vocal features like speech rate, fluency, silent pauses (especially longer than two seconds), and voice quality may mark more fine-grained cognitive alterations that could indicate preclinical AD (König et al., 2015; Szatlóczki et al., 2015; Jonell et al., 2021; Yuan et al., 2020; Roark et al., 2011a). For instance, MCI patients have been found to have a weaker and breathier voice than cognitively healthy subjects (Themistocleous et al., 2020). Categorizing words from participants' narratives into five broad categories (linguistic processes, personal concerns, psychological processes, relativity, and spoken categories) using the Linguistic Inquiry and Word Count database (Boyd et al., 2022, LIWC) has revealed that words dealing with time and space (relativity) are more sensitive to MCI detection than words from other categories (Asgari et al., 2017). Interaction patterns between interviewers and subjects during semi-structured interviews show that conversation tempo also presents distinguishing signals for detecting AD (Farzana et al., 2020; Nasreen et al., 2021; Farzana and Parde, 2022). We automatically extract diverse language and speech features from SLaCAD and identify signals and patterns from this data that may indicate early signs of AD biomarker positivity using machine learning.



SLaCAD

INV: yes but i'm gonna ask you few more questions. okay alright. can you describe when you became the leader of the dining hall please?
PAR: can i describe what?
INV: that specific day when
PAR: oh that specific day, let's see, it was my it was uh it was in my second year so that would be nineteen forty two in in this in the in the fall of forty two i would say september i can't give you the specific day but i i…

Figure 1: Characteristic language sample from SLaCAD. **INV**=Interviewer, **PAR**=Participant.

3. Approach

3.1. Data Collection

Participants were older adult volunteers (all White except one Asian participant) from a longitudinal study carried out by the University of California, San Diego's (UCSD) Shiley Marcos Alzheimer's Disease Research Center (ADRC). Exclusion criteria included those with moderate or severe AD dementia whose ability is compromised to successfully complete the task according to instructions, and those with dementia of other pathological types. ADRC participants in this study receive annual clinical and medical history, medical, neurological and neuropsychological assessments, and laboratory tests. Based on each annual evaluation, a consensus conference of neurologists and neuropsychologists determines a research diagnosis reflecting overall cognitive function (normal, MCI, or AD) based on standard diagnostic criteria (McKhann et al., 2011a). A subset of the participants (n=63) also provided CSF, via lumbar puncture, and/or blood (n=77), which was assayed for AD biomarker levels. Blood was collected within a year of the language and cognitive evaluations and CSF were collected within 5 years of the language. The research protocol was reviewed and approved by the human subject review board at UCSD and informed consent was obtained from all patients or their caregivers consistent with state law.

Language Task. Free speech samples were collected using an autobiographical interview added to the standard ADRC neuropsychological test battery. Data collection took place from 2020-2021 during the COVID pandemic and, as such, was conducted via Zoom. No specific microphone requirements were imposed but the interviewer did not proceed with the task unless the participant could be heard clearly. The participant and interviewer were recorded on the same channel. Participants were instructed to describe for five minutes a memorable event from a specific time and place during early adulthood (age 18-30). Interviewers were provided with prompts to assist participants with generating the free speech data if participants stopped talking before five minutes had passed.

	CN *n=82*	MCI *n=6*	Mild AD *n=3*
Age	75.94 (5.81)	73.16 (5.53)	74.66 (4.16)
Education	17.46 (2.13)	16.33 (1.50)	16.67 (2.31)
Sex (F/M)	44/38	1/5	1/2
Time	5.23 (2.13)	3.74 (1.24)	2.99 (0.98)
T-MoCA	19.70 (1.92)	16.33 (1.37)	13.66 (6.66)
tTau/Aβ_{42}	0.60 (1.48)	0.64 (0.50)	0.42 (0.30)
Aβ_{42}/Aβ_{40}	0.08 (0.02)	0.06 (0.03)	0.04 (0.01)
tTau	320.68 (160.33)	365.0 (195.54)	722.0 (150.33)
pTau$_{181}$	4.49 (3.06)	4.85 (1.61)	8.35 (6.19)

Table 1: Descriptive characteristics for the full dataset. Averages are reported, with standard deviations in parentheses. Time, in minutes, refers to average recording time. T-MoCA (Chappelle et al., 2023) stands for the *Telephone Montreal Cognitive Assessment* which has been administered by telephone. It uses a 22-point scale assessing auditory attention, mental flexibility, verbal fluency, sentence repetition, word-list memory, and orientation to time and place. Education is reported in years. **CN**=Cognitively Normal.

3.2. Data Transcription

All audio recordings were first automatically transcribed using the `Vosk` open-source speech recognition toolkit[2] and then the resulting transcripts were manually edited by seven undergraduate research volunteers. They were instructed to:

- Edit the transcript as needed to fix any mistakes and ensure that the text accurately matched what was said in the audio file.
- Add any missing punctuation.
- Denote words or phrases that were inaudible or questionable using the token: *(X)*.
- Add tags indicating the speaker (i.e., *Participant* or *Interviewer*).
- Add timestamps to the beginning and end of the task.
- Add tags indicating nonverbal gestures (e.g., laughs or coughs).

To ensure participant anonymity, all transcriptions were done without adding information that would compromise the identity or confidentiality of subjects. All participants were issued a unique database ID number, and all subsequent references to participants were made using only their ID number. Personnel directly associated with this project have access to the original data sheets.

3.3. Preprocessing

We preprocessed the transcripts and audio files prior to intake into the classification pipeline. The audio files, originally in .mp3 format, were converted to .wav format (44.1 kHz sample rate and 16 bits per sample). As the transcripts were segmented according to the speaker's turns,[3] we automatically added fine-grained timestamps indicating the start and end of each speaker turn. We used a forced alignment tool[4] based on the Wav2Vec2 (Baevski et al., 2020) model to generate the turn-taking timestamps. We also preprocessed the transcripts to remove interviewer utterances and speaker tags, as well as other added transcription artifacts (e.g., nonverbal cues, coughs, or laughter).

4. Corpus Analysis

SLaCAD includes autobiographical interview recordings (not available for request) and paired transcripts (available for request) for all participants, with an average task duration of 5.24 minutes (standard deviation: 1.41 minutes). We provide demographic, cognitive status, and early biomarker-related statistics in Tables 1 and 2 across different participant classes. We observe interesting patterns (in Table 1) from these descriptive statistics; for instance, cognitively normal participants clearly narrate for longer time duration than MCI and mild AD participants.

In Table 3, we provide speaker-wise statistics regarding transcript length in number of tokens, number of turns, turn length (in tokens), and turn duration for participants and interviewers. As shown, participants have a more pronounced share of the recordings than interviewers. Interviewers mostly gave task instructions, probed participants for more narrative content if they stopped talking too early, and answered clarifying questions from the participants (see the example conversation snippet in Figure 1).

[2] `https://alphacephei.com/vosk/`

[3] A *turn* is an individual speech act, defined as the full duration of time for which a single speaker is talking.

[4] `https://pytorch.org/audio/stable/tutorials/forced_alignment_tutorial.html`

	Plasma		CSF	
	pTau_{181}-	pTau_{181}+	CSF-	CSF+
	n=50	*n=27*	*n=51*	*n=12*
Age	74.46 (5.16)	78.44 (5.89)	74.22 (4.63)	77.66 (7.13)
Education	17.10 (2.37)	17.93 (1.69)	17.56 (2.05)	17.42 (2.84)
Sex (F/M)	30/20	7/20	23/28	6/6
Time	5.02 (1.96)	5.23 (1.97)	5.16 (1.96)	5.00 (1.34)
T-MoCA	19.7 (2.31)	18.44 (2.81)	19.43 (1.98)	18.91 (1.62)
tTau/Aβ_{42}	0.35 (0.30)	1.72 (3.27)	0.30 (0.12)	1.91 (2.95)
Aβ_{42}/Aβ_{40}	0.08 (0.02)	0.05 (0.03)	0.08 (0.02)	0.04 (0.01)
tTau	305.83 (159.55)	408.45 (185.37)	273.57 (106.55)	576.5 (168.34)
pTau_{181}	2.86 (0.68)	7.98 (3.39)	4.09 (2.84)	7.51 (3.49)

Table 2: Descriptive characteristics for transcripts with biomarker data. Averages are reported, with standard deviations in parentheses. Time, in minutes, refers to average recording time. T-MoCA (Chappelle et al., 2023) stands for Montreal Cognitive Assessment which has been administered by telephone. The **Plasma** column represents the 77 participants with a valid plasma (pTau_{181}) biomarker, where one subgroup is pTau_{181} negative and the other is pTau_{181} positive (Preclinical AD). The **CSF** column represents the 63 participants with valid CSF biomarkers (e.g., tTau, Aβ_{42}, or Aβ_{40}), where one subgroup is CSF (tTau and Aβ_{42} ratio) negative and the other is CSF positive (indicating preclinical AD).

5. Early AD Detection Model

To validate our dataset and assess its feasibility for relating automatically extracted language features to CSF and plasma biomarker levels, we performed preliminary experiments geared toward early AD detection. All experiments revolved around building explainable models that predict positivity for our AD biomarkers based on established cut-points (Chappelle et al., 2022):

1. **CSF tTau/Aβ_{42} Positivity** (**tTau/Aβ_{42}**): A binary variable reflecting positive versus negative status of the tTau to Aβ_{42} ratio . The positivity cutoff threshold was ≥0.609.
2. **CSF Aβ_{42}/Aβ_{40} Positivity** (**Aβ_{42}/Aβ_{40}**): A binary variable reflecting positive versus negative status of the Aβ_{42} to Aβ_{40} ratio. The positivity cutoff threshold was ≤0.056.
3. **Plasma pTau_{181} Positivity** (**pTau_{181}**): A binary variable reflecting positive versus negative status of the plasma pTau_{181} biomarker. The positivity cutoff threshold was $\geq$4.09 pg/mL.

5.1. Features

We extracted a variety of lexicosyntactic, semantic, and acoustic features from the transcripts, summarized below. All features were calculated using the participant's utterances or speech segments.

Part-Of-Speech (POS) Tags. POS tags have proven useful for detecting dementia (Masrani, 2018) and forms of primary progressive aphasia (Balagopalan et al., 2020b). We use the spaCy[5] core English POS tagger to capture the frequency of 12 coarse-grained universal POS labels (Petrov et al., 2012). Frequency counts are normalized by the number of words in the transcript.

Measure	Speaker	
	PAR	INV
Tokens	733.63±273.42	38.27±63.05
# Turns	4.70±4.86	4.09±5.03
Turn Length	374.89±348.33	6.90±7.43
Turn Duration	2.58±2.31	0.04±0.05

Table 3: Descriptive language statistics from SLaCAD, averaged across all transcripts. **INV**=Interviewer, **PAR**=Participant.

CFG Features. Context-Free Grammar (CFG) features count how often phrase structure rules (e.g., $NP \rightarrow VP\ PP$) occur in utterance parse trees, normalized by the number of nodes in the tree. CFG features have previously shown success for dementia detection (Masrani, 2018; Masrani et al., 2017). We extract parse trees using the Stanford parser (Qi et al., 2018), representing 12 Penn Treebank constituents (Marcus et al., 1993).

Syntactic Complexity. Measures of syntactic complexity have proven effective for predicting dementia from speech (Masrani, 2018). We represent utterance complexity using 16 features including parse tree depth, mean word length, mean sen-

[5] spacy.io

tence length, mean clause (noun or verb phrase) length, and number of clauses per sentence.

Named Entity Recognition (NER) Tags. NER features may be a useful and relatively domain-agnostic way to encode broad structural patterns, following the success of other intent-based features (Farzana and Parde, 2022). We extracted named entity labels using a `spaCy` model trained on the OntoNotes 5 corpus to produce the 10 fine-grained named entity types in the OntoNotes tagset (Pradhan et al., 2007). We included a frequency feature for each type, normalized by the total number of entities mentioned in the transcript.

Vocabulary Richness Features. Existing research has shown that measures of vocabulary richness can be leveraged to diagnose dementia (Masrani et al., 2017; Balagopalan et al., 2020a). We include six well-known lexical richness measures including type-token ratio (TTR), moving-average TTR (MATTR), mean segmental TTR (MSTTR), Maas index (Mass, 1972), the measure of textual lexical diversity (McCarthy, 2005, MTLD), and the hypergeometric distribution index (McCarthy and Jarvis, 2007, HD-D). We calculated each measure over the entire transcript using Python's `lexicalrichness` package.[6]

Semantic Features. We measure semantic similarity between consecutive utterances by calculating the cosine similarity between the utterance vectors and then recording the proportion of distances below three thresholds (0, 0.3, 0.5). We used averaged TF-IDF vectors to represent each utterance. We also recorded the minimum and average cosine distance between utterances.

Acoustic Features. Finally, prior work has found acoustic distinctions between subjects with and without dementia (Farzana and Parde, 2023; Masrani et al., 2017). We chunked the participant's speech segments from each recording using `Pydub`[7] and extracted 25 prosody features (Dehak et al., 2007; Vásquez-Correa et al., 2018) per chunk based on duration (i.e., number of voiced segments per second and standard deviation of duration of unvoiced segments), using the `DiSVoice`[8] tool.

5.2. Modeling and Experimental Setup

Class Balance. As observed in Table 2, data for all target variables was imbalanced:

- **tTau/Aβ_{42}**: Of 63 samples with tTau/Aβ_{42} ratios, 13 (21%) belonged to the positive class.
- **Aβ_{42}/Aβ_{40}**: Of this same set of samples, 21 (33%) belonged to the positive class.
- **pTau$_{181}$**: Of 77 samples with ptau$_{181}$ data, 27 (35%) belonged to the positive class.

To address this, we experimented with upsampling techniques and more complex approaches. We found that simple upsampling did not yield any significant performance improvements, and ultimately chose to use the Synthetic Minority Oversampling Technique (Chawla et al., 2002, SMOTE) since it increased prediction performance in our preliminary experiments.

Feature Selection. We extracted the 86 features described in §5.1 and then downsampled this feature set to a set of most informative features, experimenting with several approaches for this process. Our approaches ranked features based on three attributes: ANOVA F-values, mutual information (MI) values, and frequency among the most useful features obtained during multiple random forest classifier training rounds (RF). ANOVA and MI values were straightforward to compute. To implement RF, we repeatedly trained a random forest model (each time with a random 80% train and 20% test split) and collected the top 16 most predictive features for classifying the target value at each iteration. We then ranked all features based on their frequency in this set. To determine ideal feature set size, we then tested the prediction accuracy of an increasingly large ordered subset of features for each combination of target variable × feature selection technique. We nearly universally observed a drop in performance when using more than eight features.

Models. We experimented with both Random Forest (Breiman, 2001) and XGBoost (Chen and Guestrin, 2016) models to predict our target variables. We selected these models based on their generally high performance and explainability. We performed light hyperparameter tuning given the limited size of the dataset, to avoid overfitting. Specifically, we tuned the *max_depth* and *learning_rate* parameters for XGBoost, and the *criterion* and *n_estimator* parameters for Random Forest. We averaged the results of 1000 stratified 5-fold cross-validation runs across all combinations of target variable and feature subset, finding that Random Forest with *n_estimators = 200* and *criterion = gini* generally outperformed all other model and hyperparameter combinations.

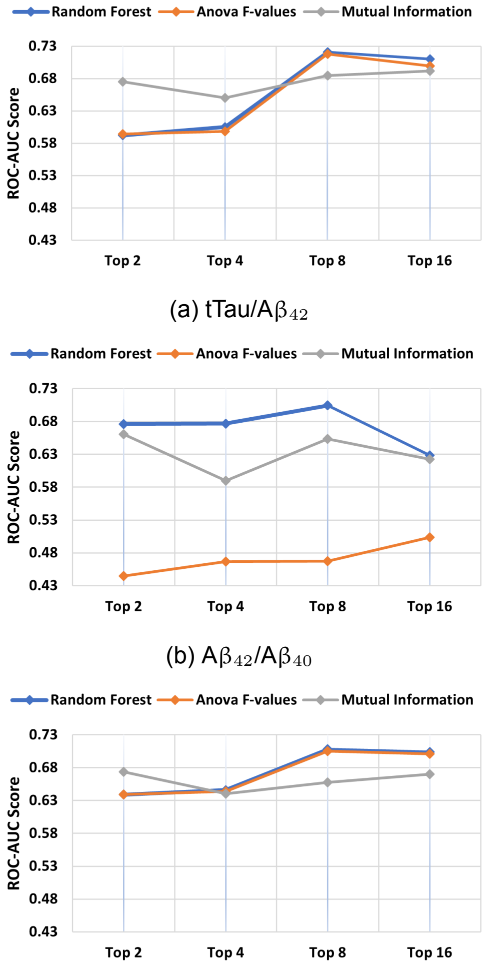
[6] `pypi.org/project/lexicalrichness`

[7] `https://pypi.org/project/pydub/`

[8] `github.com/jcvasquezc/DisVoice`

Task	Feat.	A	F_1	ROC-AUC
tTau/Aβ_{42}	2	0.67	0.30	0.59
	4	0.71	0.31	0.61
	8	**0.84**	**0.49**	**0.72**
	16	0.83	0.47	0.71
Aβ_{42}/Aβ_{40}	2	0.71	0.55	0.68
	4	0.72	0.54	0.68
	8	**0.75**	**0.58**	**0.70**
	16	0.71	0.45	0.63
pTau$_{181}$	2	0.66	0.55	0.64
	4	0.67	0.55	0.65
	8	0.73	**0.62**	**0.71**
	16	**0.74**	0.61	0.70

Table 4: Full results for the tTau/Aβ_{42}, Aβ_{42}/Aβ_{40}, and pTau$_{181}$ prediction tasks. The top n features (**Feat.**) were selected using RF for all the target variables. **A**=accuracy.



(a) tTau/Aβ_{42}

(b) Aβ_{42}/Aβ_{40}

(c) pTau$_{181}$

Figure 2: Top n features ROC-AUC score comparison for target variables.

5.3. Results

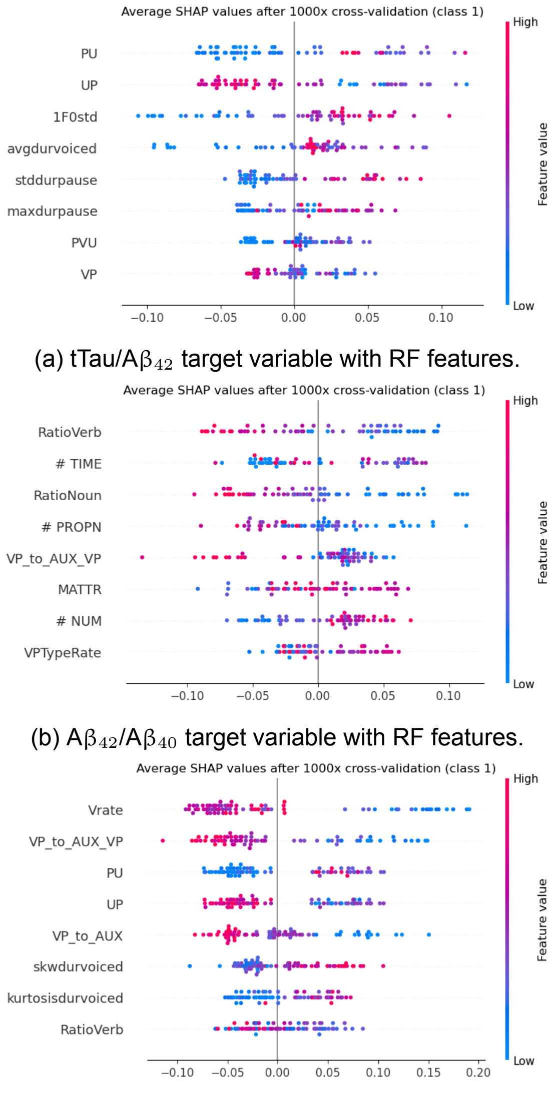
All experimental results were obtained by averaging performance across 1000 RF training/test runs with a random 80%/20% stratified split. We set n=1000 runs to ensure result stability and avoid reporting outlying values. We report performance on all target variables with an increasing number of features (top 2, top 4, top 8, and top 16) from the downsampled subsets. Although we report accuracy, F_1, and ROC-AUC, we focus on ROC-AUC since it most reliably captures performance for these tasks.

Our tTau/Aβ_{42} results are presented in Figure 2a and Table 4. We observe the highest F_1 and ROC-AUC scores using the top 8 RF features, lagging only slightly (<1.2% difference) behind the top 8 MI features. For the Aβ_{42}/Aβ_{40} results (Figure 2b and Table 4), the top 8 RF features also exhibit the best performance, with all metrics registering their highest values with this feature subset.

For the pTau$_{181}$ task (Figure 2c and Table 4), we observe that RF and ANOVA F feature selection results in very similar outcomes, with RF feature selection performing slightly better. Using the top 8 features produces the highest F_1 and ROC-AUC scores while using the top 16 features results in slightly (<1.5% difference) higher accuracy. For our ensuing feature analyses, we focus on the top 8 RF features since they are more interpretable than other feature subsets and generally exhibit the best performance across target variables.

Confounding Variables. The target variable groups were not balanced for age, sex, or years of education (Table 2). To explore potential confounding on classification results, selected features for classifying each target variable were used as input in a linear regression to predict age and education, and a linear Support Vector Classifier (SVC) to classify sex. We present the confounding variable analysis in Table 5 for each target variable. When predicting age and education via linear regression using the top 8 selected features for the corresponding target variable, we observe negative r^2 values,[9] showing that the input features failed to predict those variables. Balanced accuracies for classification of sex are slightly greater than chance, except for tTau/A$\boldsymbol{\beta_{42}}$ (for which balanced accuracy is same as chance); however, the male/female split included target variable negative and target variable positive participants in both groups. Additional details regarding the association of selected features with age, education, and sex are in Figures 4–12 in appendix A.

[9] Negative r^2 values indicate that predicting the mean dependent variable for each instance would explain more variance than a model based on the input feature.



(a) tTau/Aβ_{42} target variable with RF features.

(b) Aβ_{42}/Aβ_{40} target variable with RF features.

(c) pTau$_{181}$ target variable with RF features.

Figure 3: Shap values for the top eight features identified using mutual information (MI) or RF techniques. Details of selected features are in the appendix (Table 6 and 7)

Target Variable	Age (r^2)	Edu. (r^2)	Sex (Acc.)
tTau/Aβ_{42}	-0.7751	-0.4719	0.5098
Aβ_{42}/Aβ_{40}	-0.5057	-0.5700	0.6820
pTau$_{181}$	-0.5863	-0.5421	0.7167

Table 5: Confounding variables analysis.

Comparison with Cognitive Tests. We next compared the ability of our model to predict AD biomarker positivity compared to standard cognitive test scores to better understand whether our model may be more effective in predicting AD biomarker status compared to our current tools. More specifically, we investigated the following cognitive tests assessing global cognition: the telephone MoCA, verbal learning and memory (Craft Story Recall), attention and executive function (Oral Trail Making Parts A and B), naming and language (Animal Fluency, Letter Fluency), and working memory (Number Span). We used the scores of these tests to train a Random Forest model that would predict biomarker positivity status. We used leave-one-out cross-validation for each model, and repeated all tests 100 times to provide stable results. The only cognitive test approaching our own model performance was the Letter Fluency test, which predicted the activation of the tTau/A$\beta_4$2 with a comparable ROC-AUC score (less than 3% difference); additional details are provided in Figure 13 in the appendix A. However, this test fell short of our model when we considered the F$_1$ score, for which we observed a 15% decrease. All other models exhibited at least a 10% decrease across all metrics compared to our model trained on speech and language biomarkers.

5.4. Discussion

For tTau/Aβ_{42}, Aβ_{42}/Aβ_{40}, and pTau$_{181}$, RF feature selection generated the best performance in terms of ROC-AUC score. Overall, we observed better performance predicting the plasma pTau$_{181}$ target variable across all metrics than CSF target variables (tTau/Aβ_{42} and Aβ_{42}/Aβ_{40}). This may be because the collection of plasma was more proximal in time (≤1 year) to the language assessment compared to CSF collection (≤5 years). We evaluated the explainability of our models using SHAP values (Lundberg and Lee, 2017) for all target variables in figures 3a–3c, and discuss our findings below.

CSF tTau/Aβ_{42}. Although we observed higher overall tTau prediction ROC-AUC than other target variables, we also still observed (based on the difference between accuracy and F$_1$) that predictions may be biased towards the negative class. Interestingly, the most predictive features (Figure 3a) were acoustic. Furthermore, almost all of these features exhibited clear correlations with the target value. The positively correlated feature *avgdurvoiced* suggests that individuals at the AD preclinical stage may struggle to remain on topic. This is also confirmed by the negative correlation of the *VP* feature, which expresses the ratio between voiced versus paused time in a conversation. We expected this correlation since individuals with healthy cognition should have fewer pauses in their speech (Farzana and Parde, 2020).

CSF $\mathbf{A\beta_{42}/A\beta_{40}}$. The results for $A\beta_{42}/A\beta_{40}$ are moderately strong and also confirm an interesting trend seen with tTau, suggesting that the features *RatioVerb* and *RatioNoun* are negatively correlated with $A\beta_{42}/A\beta_{40}$ positivity (see Figure 3b) as cognitively impaired subjects tend to use more function words (Farzana and Parde, 2020).

Plasma $\mathbf{pTau_{181}}$. Finally, for the $pTau_{181}$ target variable we observe overall balanced metrics and the top F_1 among all tasks. It is interesting to note that most of the top features for this target variable are again audio-related, with clear correlations. We observe a highly negative correlation for the *Vrate* (voicerate meaning speaking rate) feature, which have been found to be the earliest measurable speech feature for individuals in early stages of cognitive decline (Szatlóczki et al., 2015). For the *VP_TO_AUX_VP* feature, we observe the same strong negative correlation as observed for the $A\beta_{42}/A\beta_{40}$ target variable. Another interesting correlation is the positive correlation for the *kurtosisdurvoiced* feature, which means more inconsistent speech duration distribution from subjects who are $pTau_{181}$ positive compared to subjects who are $pTau_{181}$ negative.

Common Features. Between $A\beta_{42}/A\beta_{40}$ and $pTau_{181}$ target variables, linguistic (RatioVerb), and syntactic (VP_to_AUX_VP) features are negatively correlated showing that less use of function words are commonly observed in those who were $A\beta_{42}/A\beta_{40}$ and $pTau_{181}$ positive. In contrast, we observe that $pTau_{181}$ and $tTau/A\beta_{42}$ target variables are positively associated with the acoustic feature *PU* (the ratio of pause duration to unvoiced segment duration), meaning more pauses were observed in those who were $pTau_{181}$ and $tTau/A\beta_{42}$ positive.

6. Conclusion

We present a new spoken language corpus, SLaCAD, containing spontaneous speech transcripts, derived linguistic and acoustic markers, and comprehensive cognitive characterization in 91 older adults who are predominantly cognitively normal. The sample has been divided into two groups: one group of 63 participants with CSF-related AD biomarker levels available, and another group of 77 participants with plasma-related AD biomarker levels. We detailed the data collection procedures and transcription process, and generated speech and language features from the resulting transcripts and audio recordings to build explainable models capable of detecting early AD characteristics. We found that some of the speech and language features, such as specific POS frequency and prosody features that previously proved to be effective in AD detection (Masrani et al., 2017; Farzana and Parde, 2020), also relate to our early AD target variables. Furthermore, we identified correlations between audio features and Tau-related biomarkers. Our experiments provide promising initial results on this dataset for detecting early AD using speech and language biomarkers. However, further extensive research and validation in larger samples is needed before drawing definitive conclusions or establishing clinical benchmarks for these preliminary findings. In keeping with our study and ethics protocols, SLaCAD (the transcriptions and derived linguistic and acoustic features) will be publicly available via data requests through the National Alzheimer's Coordinating Center.[10]

7. Ethics Statement

7.1. Limitations

This work is limited by several factors. In general, caution should be taken whenever computationally exploring datasets without theory-guided hypotheses, as outlined in detail by Hitczenko et al. (2020). Moreover, this work was the result of substantial effort. For instance, it took seven transcribers more than one year to fix the transcripts included in this corpus, with initial experiments using only automated speech recognizers failing to independently produce workable transcripts. Thus, although using automatically extracted language features to predict preclinical AD appears feasible or at least promising from our preliminary evidence, there is still much work to be done before this process could be reasonably used as a replacement for CSF collection. Finally, datasets within the cognitive health domain are notoriously small (Farzana and Parde, 2023). It is difficult to make strong claims given our sample size, and the lack of racial and ethnic diversity in the study cohort make it unclear whether our findings would generalize to broader or differently-distributed subject populations. Collectively, these limitations offer substantial potential for future research growth within this crucial domain.

7.2. Potential Risks

This dataset includes real-world language samples collected from individuals, paired with labels

[10] Contact any of the authors for dataset access.

indicating their Alzheimer's disease status and CSF/blood plasma biomarker levels. Careful steps were taken to anonymize this data and handle it responsibly and respectfully, in accordance with our approved IRB protocol. Although we do not anticipate this occurring, if participants' identities became public this information could be compromising since AD status is considered a sensitive or private topic by many.

Moreover, use of this dataset as intended may lead to meaningful clinical discovery regarding language and its association with AD pathology. It could also lead to the development of models that automatically predict pre-clinical AD status. An *intended use* of such a model would be to support trained clinicians by helping to quickly identify patients at early stages who many need further review. An *unintended use* of such a model would be to act as a replacement for clinical professionals, or to trust its judgment without further review.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback, and our seven undergraduate transcribers for their hard work towards making this dataset possible. The creation of SLaCAD was funded in part by a seed grant from the University of California San Diego's Alzheimer's Disease Research Center. S. Farzana and N. Parde were also partially funded during this time by the National Science Foundation under Grant No. 2125411. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. Bibliographical References

Samrah Ahmed, Anne-Marie F Haigh, Celeste A de Jager, and Peter Garrard. 2013. Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. *Brain : a journal of neurology*, 136(Pt 12):3727—3737.

Marilyn S. Albert, Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C. Fox, Anthony Gamst, David M. Holtzman, William J. Jagust, Ronald C. Petersen, Peter J. Snyder, Maria C. Carrillo, Bill Thies, and Creighton H. Phelps. 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. *Alzheimer's & Dementia*, 7(3):270–279.

N Andreasen, L Minthon, A Clarberg, P Davidsson, J Gottfries, E Vanmechelen, H Vanderstichele, B Winblad, and K Blennow. 1999. Sensitivity, specificity, and stability of csf-tau in ad in a community-based patient sample. *Neurology*, 53(7):1488—1494.

Niels Andreasen, Eugeen Vanmechelen, Hugo Vanderstichele, Pia Davidsson, and Kaj Blennow. 2003. Cerebrospinal fluid levels of total-tau, phospho-tau and aβ42 predicts development of alzheimer's disease in patients with mild cognitive impairment. *Acta Neurologica Scandinavica*, 107(s179):47–51.

Hiroyuki Arai, Koichi Ishiguro, Hideto Ohno, Michiko Moriyama, Nobuo Itoh, Nobuyuki Okamura, Toshifumi Matsui, Yu ichi Morikawa, Etsuo Horikawa, Hideki Kohno, Hidetada Sasaki, and Kazutomo Imahori. 2000. Csf phosphorylated tau protein and mild cognitive impairment: A prospective study. *Experimental Neurology*, 166(1):201–203.

Eider M. Arenaza-Urquijo and Prashanthi Vemuri. 2018. Resistance vs resilience to alzheimer disease. *Neurology*, 90:695 – 703.

Meysam Asgari, Jeffrey Kaye, and Hiroko Dodge. 2017. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020a. To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection. In *Proc. Interspeech 2020*, pages 2167–2171.

Aparna Balagopalan, Jekaterina Novikova, Matthew B A Mcdermott, Bret Nestor, Tristan Naumann, and Marzyeh Ghassemi. 2020b. Cross-Language Aphasia Detection using Optimal Transport Domain Adaptation. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 202–219. PMLR.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994.

The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*.

Kaj Blennow, Mony J de Leon, and Henrik Zetterberg. 2006. Alzheimer's disease. *The Lancet*, 368(9533):387–403.

Kaj Blennow and Harald Hampel. 2003. Csf markers for incipient alzheimer's disease. *The Lancet Neurology*, 2(10):605–613.

Ryan Boyd, Ashwini Ashokkumar, Sarah Seraj, and James Pennebaker. 2022. The development and psychometric properties of liwc-22.

L Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Jony A. Pinto Jr Breno S. O. Diniz and Orestes Vicente Forlenza. 2008. Do csf total tau, phosphorylated tau, and β-amyloid 42 help to predict progression of mild cognitive impairment to alzheimer's disease? a systematic review and meta-analysis of the literature. *The World Journal of Biological Psychiatry*, 9(3):172–182. PMID: 17886169.

Marc Brysbaert and Boris New. 2009. Moving beyond kucera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41:977–90.

Marco Canevelli, Nawal Adali, Cécile Tainturier, Giuseppe Bruno, Matteo Cesari, and Bruno Vellas. 2013. Cognitive interventions targeting subjective cognitive complaints. *American Journal of Alzheimer's Disease & Other Dementias®*, 28(6):560–567. PMID: 23823142.

Sheridan Chappelle, Christina Gigliotti, Gabriel Leger, Guerry Peavy, Diane Jacobs, Sarah Banks, Emily Little, Douglas Galasko, and David Salmon. 2022. Comparison of the telephone□montreal cognitive assessment (t□moca) and telephone interview for cognitive status (tics) as remote screening tests for early alzheimer's disease. *Alzheimer's & Dementia*, 18.

Sheridan D. Chappelle, Christina Gigliotti, Gabriel C. Léger, Guerry M. Peavy, Diane M. Jacobs, Sarah J. Banks, Emily A. Little, Douglas Galasko, and David P. Salmon. 2023. Comparison of the telephone-montreal cognitive assessment (t-moca) and telephone interview for cognitive status (tics) as screening tests for early alzheimer's disease. *Alzheimer's & Dementia*, 19(10):4599–4608.

Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Natasha Clarke, Thomas Barrick, and Peter Garrard. 2021. A comparison of connected speech tasks for detecting early alzheimer's disease and mild cognitive impairment using natural language processing and machine learning. *Frontiers in Computer Science*, 3.

Najim Dehak, Pierre Dumouchel, and Patrick Kenny. 2007. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2095–2103.

Ira Driscoll and Juan Troncoso. 2011. Asymptomatic alzheimers disease: A prodrome or a state of resilience? *Current Alzheimer research*, 8:330–5.

Bruno Dubois, Harald Hampel, Howard H. Feldman, Philip Scheltens, Paul Aisen, Sandrine Andrieu, Hovagim Bakardjian, Habib Benali, Lars Bertram, Kaj Blennow, Karl Broich, Enrica Cavedo, Sebastian Crutch, Jean-François Dartigues, Charles Duyckaerts, Stéphane Epelbaum, Giovanni B. Frisoni, Serge Gauthier, Remy Genthon, Alida A. Gouw, Marie-Odile Habert, David M. Holtzman, Miia Kivipelto, Simone Lista, José-Luis Molinuevo, Sid E. O'Bryant, Gil D. Rabinovici, Christopher Rowe, Stephen Salloway, Lon S. Schneider, Reisa Sperling, Marc Teichmann, Maria C. Carrillo, Jeffrey Cummings, and Cliff R. Jack. 2016. Preclinical alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*, 12(3):292–323.

Shahla Farzana and Natalie Parde. 2020. Exploring MMSE Score Prediction Using Verbal and Non-Verbal Cues. In *Proc. Interspeech 2020*, pages 2207–2211.

Shahla Farzana and Natalie Parde. 2022. Are interaction patterns helpful for task-agnostic dementia detection? an empirical exploration. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 172–182, Edinburgh, UK. Association for Computational Linguistics.

Shahla Farzana and Natalie Parde. 2023. Towards domain-agnostic and domain-adaptive dementia detection from spoken language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11965–11978, Toronto, Canada. Association for Computational Linguistics.

Shahla Farzana, Mina Valizadeh, and Natalie Parde. 2020. Modeling dialogue in conversational cognitive health screening interviews. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Daniel Ferreira, Amado Rivero-Santana, Lilisbeth Perestelo-Pérez, Eric Westman, Lars-Olof Wahlund, Antonio Sarría, and Pedro Serrano-Aguilar. 2014. Improving csf biomarkers' performance for predicting progression from mild cognitive impairment to alzheimer's disease by considering different confounding factors: A meta-analysis. *Frontiers in Aging Neuroscience*, 6.

Katrina Forbes-McKay and A Venneri. 2005a. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, 26:243–54.

KE Forbes-McKay and A Venneri. 2005b. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, 26(4):243—254.

D. Galasko, C. Clark, L. Chang, B. Miller, R. C. Green, R. Motter, and P. Seubert. 1997. Assessment of csf levels of tau protein in mildly demented patients with alzheimer's disease. *Neurology*, 48(3):632–635.

Peter Garrard, Lisa M. Maloney, John R. Hodges, and Karalyn Patterson. 2004a. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128(2):250–260.

Peter Garrard, Lisa M. Maloney, John R. Hodges, and Karalyn Patterson. 2004b. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128(2):250–260.

Johan Gottfries, Kaj Blennow, Martin Lehmann, Björn Regland, and Carl-Gerhard Gottfries. 2001. One-carbon metabolism and other biochemical correlates of cognitive impairment as visualized by principal component analysis. *Journal of geriatric psychiatry and neurology*, 14:109–14.

H. RANDALL GRIFFITH, KELLI L. NETSON, LINDY E. HARRELL, EDWARD Y. ZAMRINI, JOHN C. BROCKINGTON, and DANIEL C. MARSON. 2006. Amnestic mild cognitive impairment: Diagnostic outcomes and clinical prediction over a two-year time period. *Journal of the International Neuropsychological Society*, 12(2):166–175.

Ihab Hajjar, Maureen Okafor, Jinho Choi, Elliot Moore, Anees Abrol, Vince Calhoun, and Felicia Goldstein. 2023. Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 15.

Kasia Hitczenko, Vijay A Mittal, and Matthew Goldrick. 2020. Understanding Language Abnormalities and Associated Clinical Markers in Psychosis: The Promise of Computational Methods. *Schizophrenia Bulletin*, 47(2):344–362.

Krister Håkansson, Tiia Ngandu, and Miia Kivipelto. 2018. The patient with cognitive impairment. In *Treatable and Potentially Preventable Dementias*, page 52–80. Cambridge University Press.

Koichi Ishiguro, Hideto Ohno, Hiroyuki Arai, Haruyasu Yamaguchi, Katsuya Urakami, Jung-Mi Park, Kazuki Sato, Hideki Kohno, and Kazutomo Imahori. 1999. Phosphorylated tau in human cerebrospinal fluid is a diagnostic marker for alzheimer's disease. *Neuroscience Letters*, 270(2):91–94.

Clifford R. Jack Jr., David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, Reisa Sperling, Contributors, Cerise Elliott, Eliezer Masliah, Laurie Ryan, and Nina Silverberg. 2018. Nia-aa research framework: Toward a biological definition of alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562.

Shorena Janelidze, Niklas Mattsson, Sebastian Palmqvist, Ruben Smith, Thomas G Beach, Geidy E Serrano, Xiyun Chai, Nicholas K

Proctor, Udo Eichenlaub, Henrik Zetterberg, Kaj Blennow, Eric M Reiman, Erik Stomrud, Jeffrey L Dage, and Oskar Hansson. 2020. Plasma p-tau181 in alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to alzheimer's dementia. *Nature medicine*, 26(3):379â€"386.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Patrik Jonell, Birger Moëll, Krister Håkansson, Gustav Eje Henter, Taras Kucherenko, Olga Mikheeva, Göran Hagman, Jasper Holleman, Miia Kivipelto, Hedvig Kjellström, Joakim Gustafson, and Jonas Beskow. 2021. Multimodal capture of patient behaviour for improved detection of early dementia: Clinical feasibility and preliminary results. *Frontiers in Computer Science*, 3.

Miia Kivipelto, Francesca Mangialasche, and Tiia Ngandu. 2017. Can lifestyle changes prevent cognitive impairment? *The Lancet. Neurology*, 16.

Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H. Robert, and Renaud David. 2015. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124.

Linda Lang, Angela Clifford, Li Wei, Dongmei Zhang, Daryl Leung, Glenda Augustine, Isaac Danat, Weiju Zhou, John Copeland, Kaarin Anstey, and Ruoling Chen. 2017. Prevalence and determinants of undetected dementia in the community: A systematic literature review and a meta-analysis. *BMJ Open*, 7:e011146.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Vaden Masrani. 2018. Detecting dementia from written and spoken language. Master's thesis, University of British Columbia.

Vaden Masrani, Gabriel Murray, Thalia Shoshana Field, and Giuseppe Carenini. 2017. Domain adaptation for detecting mild cognitive impairment. In *Advances in Artificial Intelligence*, pages 248–259, Cham. Springer International Publishing.

Heinz-Dieter Mass. 1972. Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 2(8):73.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M. McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack, Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, Richard C. Mohs, John C. Morris, Martin N. Rossor, Philip Scheltens, Maria C. Carrillo, Bill Thies, Sandra Weintraub, and Creighton H. Phelps. 2011a. 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. *Alzheimer's & Dementia*, 7(3):263–269.

Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack Jr., Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, Richard C. Mohs, John C. Morris, Martin N. Rossor, Philip Scheltens, Maria C. Carrillo, Bill Thies, Sandra Weintraub, and Creighton H. Phelps. 2011b. 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. *Alzheimer's & Dementia*, 7(3):263–269.

Jonilda Mecollari Mueller, Bruce Hermann and Lyn S. Turkstra. 2018. Connected speech and language in mild cognitive impairment and alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology*, 40(9):917–939. PMID: 29669461.

Kimberly Mueller, Carol Van Hulle, Rebecca Langhough, Erin Jonaitis, Cassandra Peters, Tobey Betthauser, Bradley Christian, Nathaniel

Chin, Bruce Hermann, and Sterling Johnson. 2021. Amyloid beta associations with connected speech in cognitively unimpaired adults. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 13.

Shamila Nasreen, Julian Hough, and Matthew Purver. 2021. Rare-class dialogue act tagging for Alzheimer's disease diagnosis. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 290–300, Singapore and Online. Association for Computational Linguistics.

Peter Nestor and Ph Scheltens. 2004. Advances in the early detection of alzheimer's disease (review). *Nature medicine*, 10 Suppl:S34–41.

Matthew P. Pase, Alexa S. Beiser, Jayandra J. Himali, Claudia L. Satizabal, Hugo J. Aparicio, Charles DeCarli, Geneviève Chêne, Carole Dufouil, and Sudha Seshadri. 2019. Assessment of Plasma Total Tau Level as a Predictive Biomarker for Dementia and Related Endophenotypes. *JAMA Neurology*, 76(5):598–606.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *International Conference on Semantic Computing (ICSC 2007)*, pages 517–526.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

B. Roark, M. Mitchell, J. Hosom, K. Hollingshead, and J. Kaye. 2011a. Spoken language derived measures for detecting mild cognitive impairment. *Trans. Audio, Speech and Lang. Proc.*, 19(7):2081–2090.

B. Roark, M. Mitchell, J. Hosom, K. Hollingshead, and J. Kaye. 2011b. Spoken language derived measures for detecting mild cognitive impairment. *Trans. Audio, Speech and Lang. Proc.*, 19(7):2081–2090.

José Rafael Romero, Serkalem Demissie, Alexa Beiser, Jayandra J. Himali, Charles C. DeCarli, Daniel Levy, and Sudha Seshadri. 2020. Relation of plasma β□amyloid, clusterin, and tau with cerebral microbleeds: Framingham heart study. *Annals of Clinical and Translational Neurology*, 7:1083 – 1091.

Ayda Rostamzadeh, Lara Bohr, Michael Wagner, Christopher Baethge, and Frank Jessen. 2022. Progression of subjective cognitive decline to mci or dementia in relation to biomarkers for alzheimer disease: A meta-analysis. *Neurology*, 99:10.1212/WNL.0000000000201072.

Michal Tomek Seyed Ahmad Sajjadi, Karalyn Patterson and Peter J. Nestor. 2012. Abnormalities of connected speech in semantic dementia vs alzheimer's disease. *Aphasiology*, 26(6):847–866.

Kamlesh Sharma. 2019. Cholinesterase inhibitors as alzheimer's therapeutics (review). *Molecular Medicine Reports*, 20.

Mikio Shoji, Etsuro Matsubara, Mitsuyasu Kanai, Mitsunori Watanabe, Tamiko Nakamura, Yasushi Tomidokoro, Masami Shizuka, Katsumi Wakabayashi, Yukifusa Igeta, Yoshio Ikeda, Kazuyuki Mizushima, Masakuni Amari, Koji Ishiguro, Takeshi Kawarabayashi, Yasuo Harigaya, Koich Okamoto, and Shunsaku Hirai. 1998. Combination assay of csf tau, aβ1-40 and aβ1-42(43) as a biochemical marker of alzheimer's disease. *Journal of the Neurological Sciences*, 158(2):134–140.

Antoine Slegers, Renée pier Filiou, Maxime Montembeault, and Simona Maria Brambati. 2018. Connected speech features from picture description in alzheimer's disease: A systematic review. *Journal of Alzheimer's disease : JAD*, 65 2:519–542.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

Gréta Szatlóczki, Ildikó Hoffmann, Veronika Vincze, János Kálmán, and Magdolna Pákáski. 2015. Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in Aging Neuroscience*, 7.

Gréta Szatlóczki, Ildiko Hoffmann, Veronika Vincze, János Kálmán, and Magdolna Pakaski.

2015. Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in Aging Neuroscience*, 7.

Charalambos Themistocleous, Marie Eckerström, and Dimitrios Kokkinakis. 2020. Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. *PLOS ONE*, 15:e0236009.

Sander Verfaillie, Jurriaan Witteman, Rosalinde E R Slot, Ilanah Pruis, Lieke Vermaat, Niels Prins, Niels Schiller, Mark Wiel, Philip Scheltens, Bart Berckel, Wiesje Flier, and Sietske Sikkes. 2019. High amyloid burden is associated with fewer specific words during spontaneous speech in individuals with subjective cognitive decline. *Neuropsychologia*.

Rohit Voleti, Julie M. Liss, and Visar Berisha. 2020. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):282–298.

J.C. Vásquez-Correa, J.R. Orozco-Arroyave, T. Bocklet, and E. Nöth. 2018. Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease. *Journal of Communication Disorders*, 76:21–36.

Andrew J. Wawrzyniak. 2020. Framingham heart study. In *Encyclopedia of Behavioral Medicine*, pages 1–4, New York, NY. Springer New York.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *Proceedings of INTERSPEECH*, pages 2162–2166.

Zhen Zhao, Amy R Nelson, Christer Betsholtz, and Berislav V Zlokovic. 2015. Establishment and dysfunction of the blood-brain barrier. *Cell*, 163(5):1064–1078.

A. Appendix

Category	**Selected Features**	**Definition**
POS tags	# AUX	# auxiliary verb tokens
	# VERB	# verb tokens
	#PROPN	# proper noun tokens
	# CCONJ	# conjunction tokens
	RatioVerb	percentage of tokens with verb POS tag
	RatioNoun	percentage of tokens with noun POS tag
NER tags	#DATE	# tokens associated with date
	#TIME	# tokens associated with time
	# NUM	# tokens associated with number
CFG	VP_to_AUX_ADJP	sentence structure with phrase type verb phrase to auxiliary to adjective phrase
	VP_to_AUX_VP	sentence structure with phrase type verb phrase to auxiliary to verb phrase
	VP_to_AUX	sentence structure with phrase type verb phrase to auxiliary
Syntactic Complexity	VPTypeRate	ratio of # verb phrases in parse tree of a sentence and # words in the sentence
Vocabulary Richness	# Unique Tokens	# unique tokens available in the transcript
	MATTR	moving average of type-token ratio (TTR)

Table 6: Descriptions of selected lexicosyntactic features in modeling different target variables of preclinical AD.

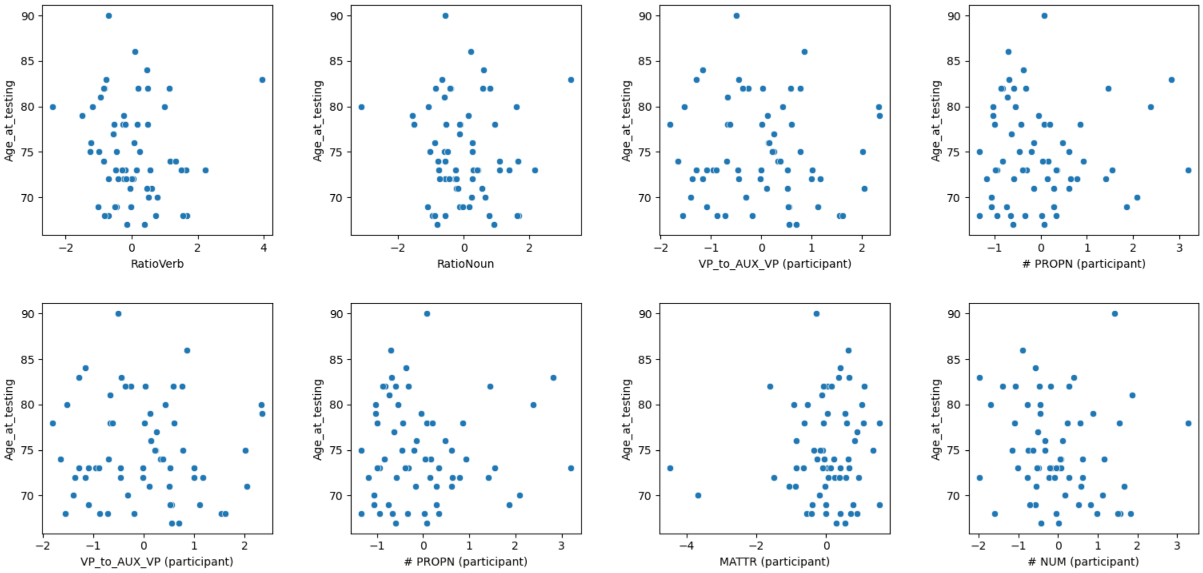


Figure 4: Association of top 8 selected features of Aβ_{42}/Aβ_{40} variable with *age*.

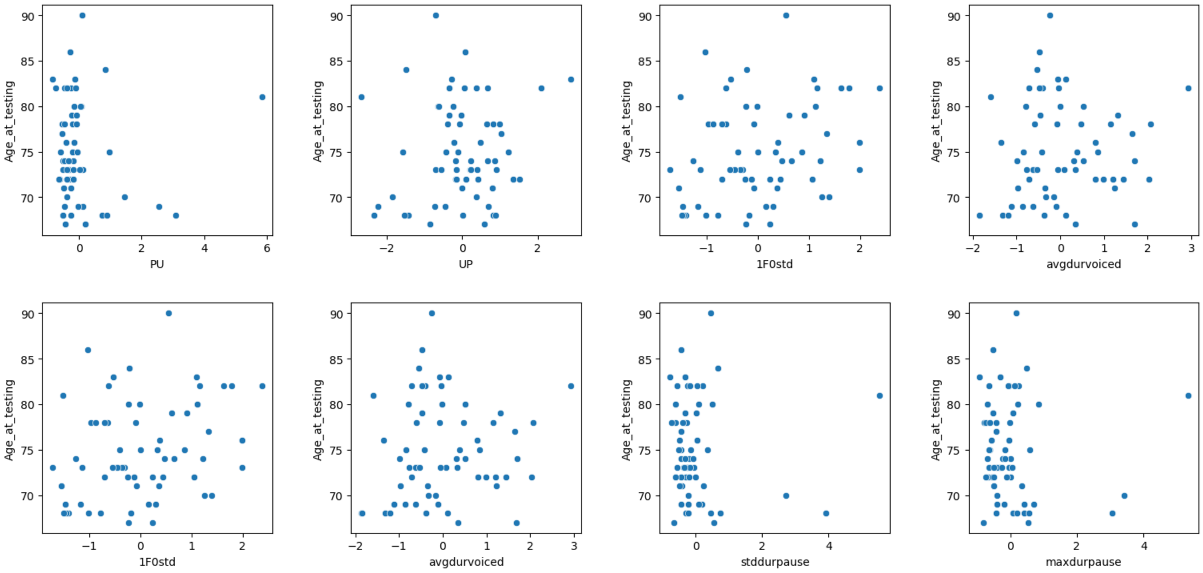


Figure 5: Association of top 8 selected features of tTau/Aβ_{42} variable with *age*.

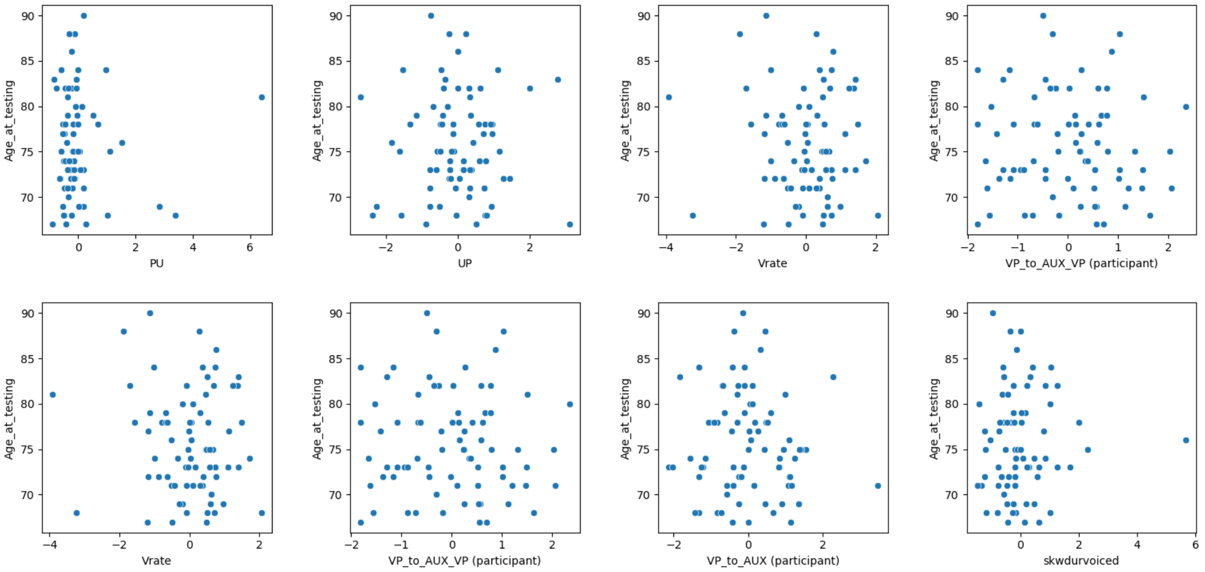


Figure 6: Association of top 8 selected features of pTau_{181} variable with *age*.

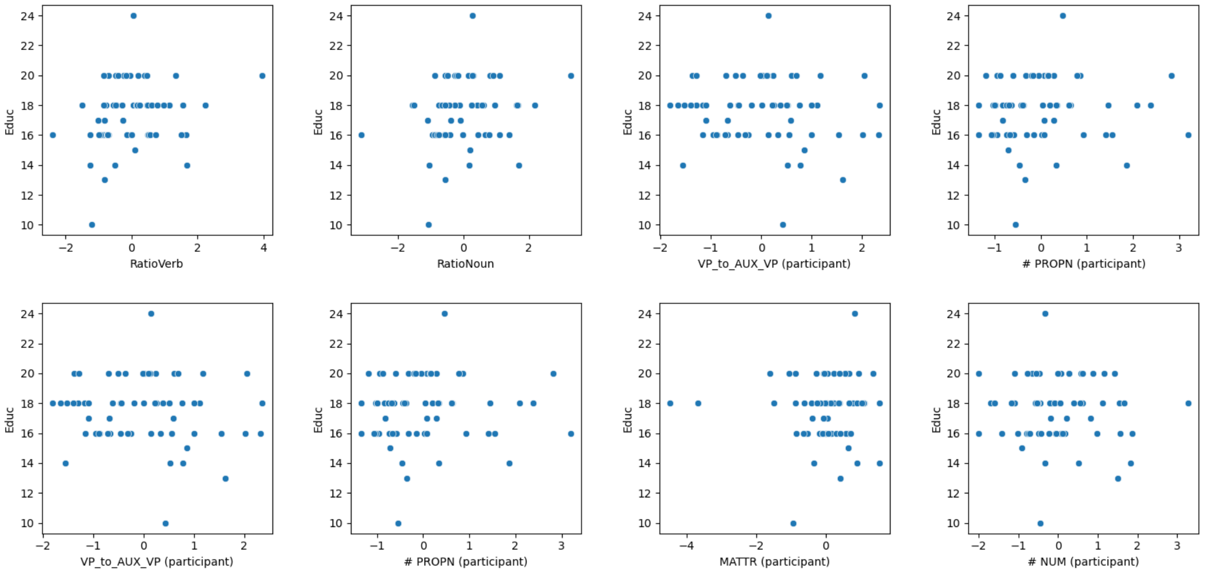


Figure 7: Association of top 8 selected features of $A\beta_{42}/A\beta_{40}$ variable with *education* (in year).

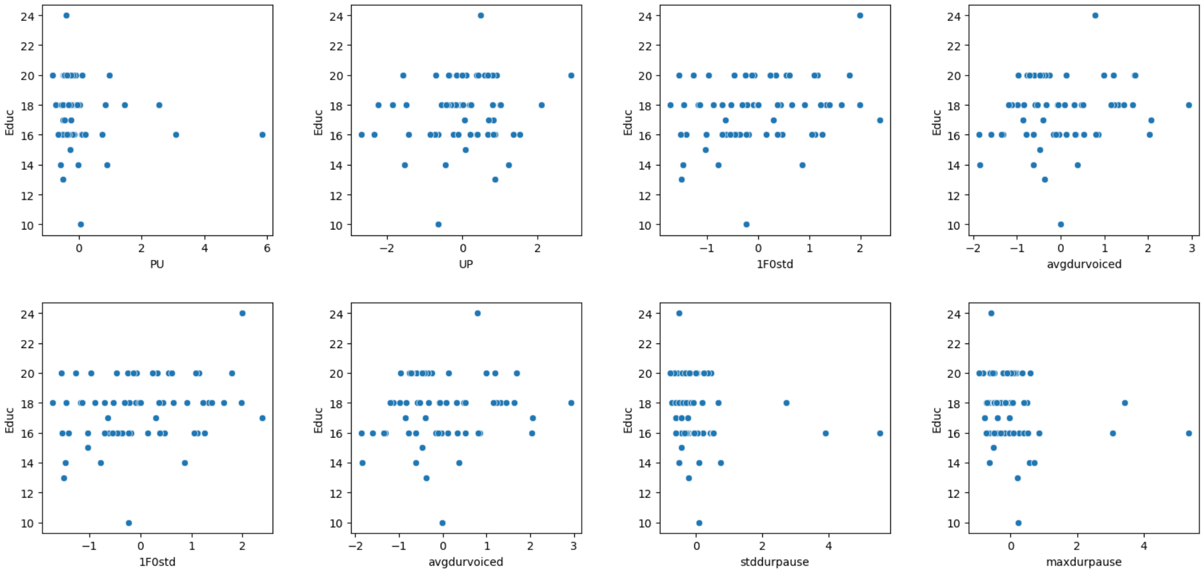


Figure 8: Association of top 8 selected features of tTau/Aβ_{42} variable with *education* (in year).

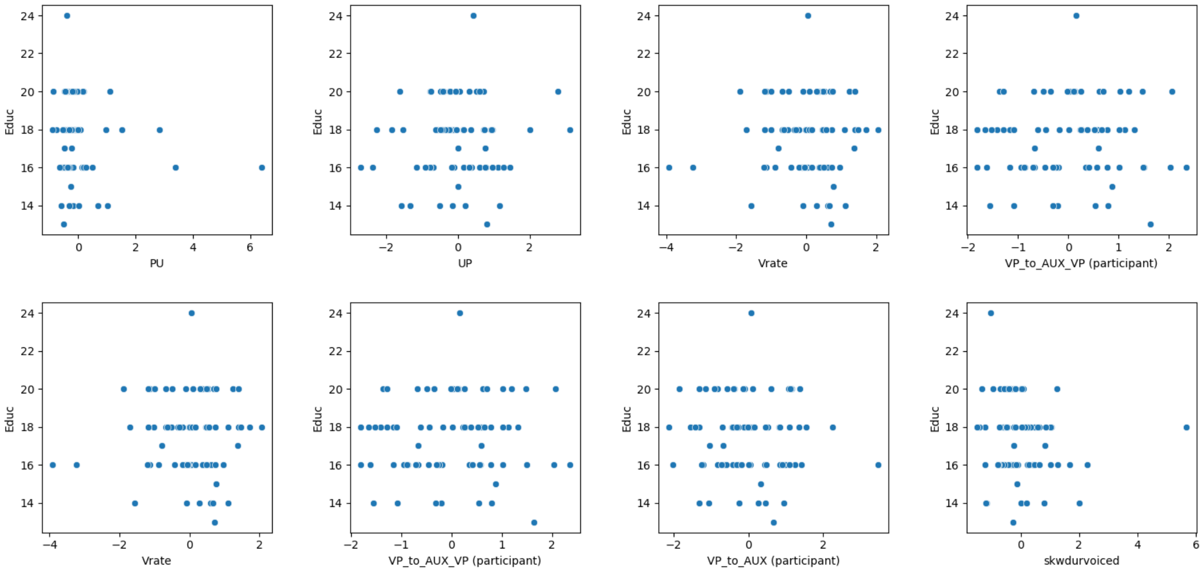


Figure 9: Association of top 8 selected features of pTau$_{181}$ variable with *education* (in year).

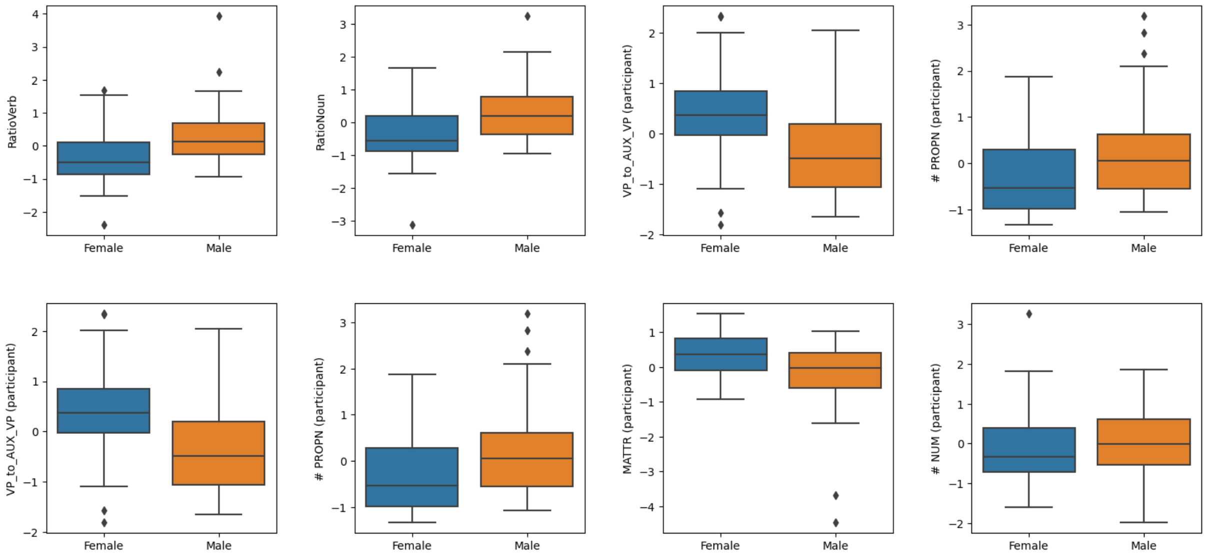


Figure 10: Association of top 8 selected features of $A\beta_{42}/A\beta_{40}$ variable with *gender*.

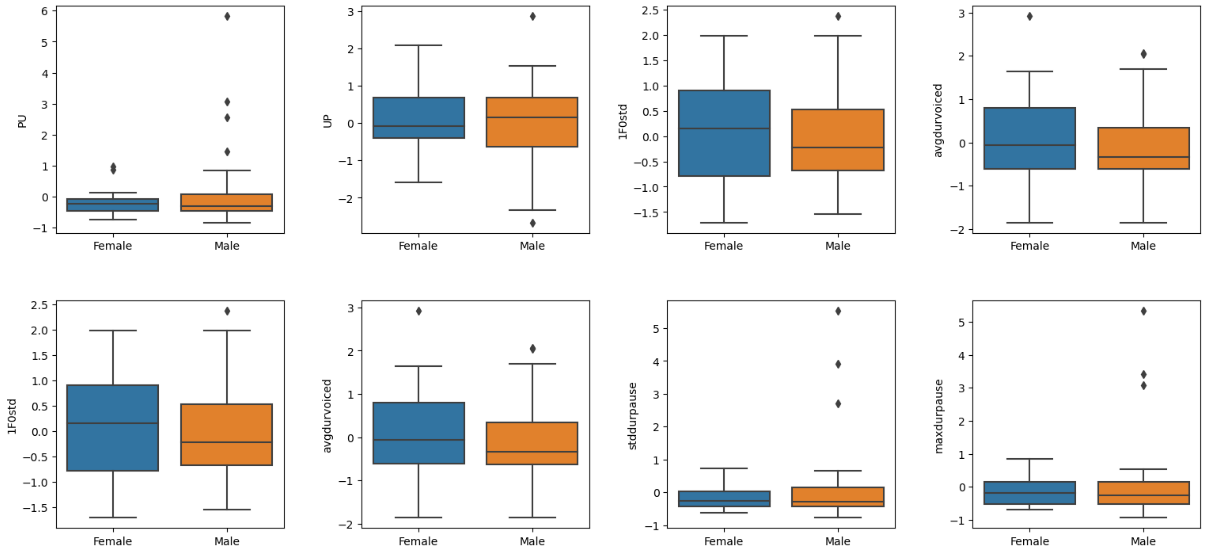


Figure 11: Association of top 8 selected features of tTau/$A\beta_{42}$ variable with *gender*.

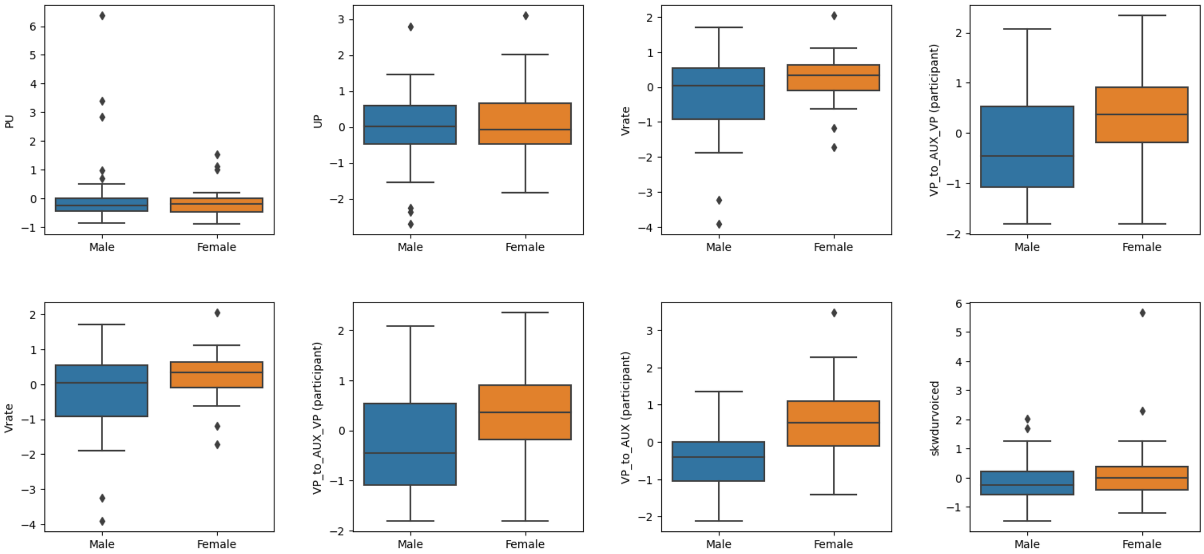


Figure 12: Association of top 8 selected features of pTau_{181} variable with *gender*.

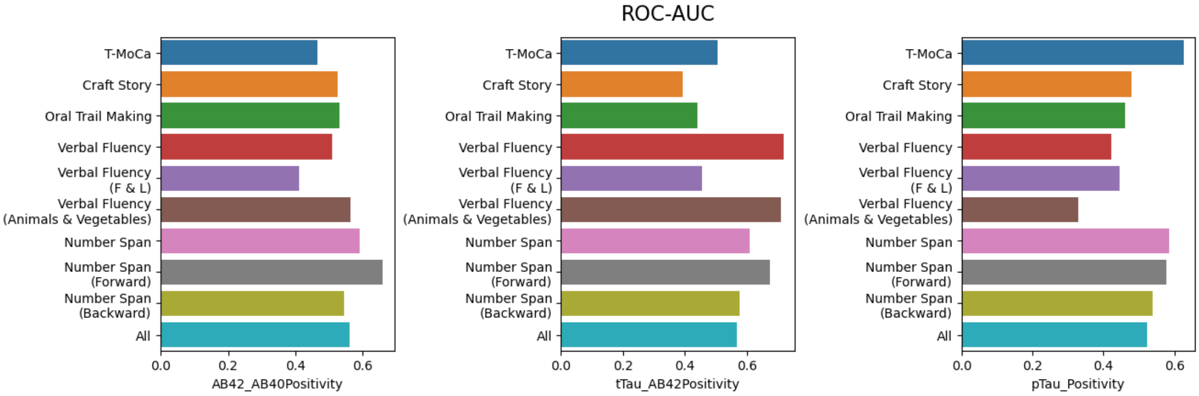


Figure 13: ROC-AUC metric of the standard cognitive tests in predicting the target variables related to preclinical AD.

Category	Selected Features	Definition
Audio	PU	pause/unvoiced
	UP	unvoiced/pause
	avgdurvoiced	average duration of voiced segment
	stddurpause	standard deviation of pause duration
	maxdurpause	maximum pause duration
	PVU	pause duration/(voiced duration+unvoiced duration)
	VP	voiced duration/pause duration
	Vrate	# voiced segments per second (voiced rate)
	skwdurvoiced	skewness of duration of voiced segments
	kurtosisdurvoiced	kurtosis of duration of voiced segments
	1F0std	standard deviation of fundamental freq. features in first voiced segment

Table 7: Descriptions of selected acoustic features in modeling different target variables of preclinical AD.