Non-Invasive Classification of Alzheimer's Disease Using Eye Tracking and Language

Oswald Barral* OBARRAL@CS.UBC.CA
Department of Computer Science, University of British Columbia
Vancouver, BC, Canada
**Alphabetical order, authors with equal contribution*

Hyeju Jang* HYEJUJ@CS.UBC.CA
Department of Computer Science, University of British Columbia
Vancouver, BC, Canada
**Alphabetical order, authors with equal contribution*

Sally Newton-Mason S.NEWTON-MASON@ALUMNI.UBC.CA
Division of Neurology, Faculty of Medicine, University of British Columbia
Vancouver, BC, Canada

Sheetal Shajan SHEETAL.SHAJAN@ALUMNI.UBC.CA
Division of Neurology, Faculty of Medicine, University of British Columbia
Vancouver, BC, Canada

Thomas Soroski TOM.SOROSKI@UBC.CA
Division of Neurology, Faculty of Medicine, University of British Columbia
Vancouver, BC, Canada

Giuseppe Carenini CARENINI@CS.UBC.CA
Department of Computer Science, University of British Columbia
Vancouver, BC, Canada

Cristina Conati CONATI@CS.UBC.CA
Department of Computer Science, University of British Columbia
Vancouver, BC, Canada

Thalia Field THALIA.FIELD@UBC.CA
Division of Neurology, Faculty of Medicine and Djavad Mowafaghian Centre for Brain Health
University of British Columbia
Vancouver, BC, Canada

Abstract

Alzheimer's disease (AD) is an insidious progressive neurodegenerative disease resulting in impaired cognition, dementia, and eventual death. At the earliest stages of the disease, decline in multiple cognitive domains including speech and eye movements occurs, and worsens with disease progression. Therefore, investigating speech and eye movements is promising as a non-invasive method for early classification of AD. While related work has investigated AD classification using speech collected during spontaneous speech tasks, no prior research has studied the utility of eye movements and their combination with speech for this classification task. In this paper, we present classification experiments with speech and eye movement data collected from 68 memory clinic patients (with a diagnosis of AD, mixed dementia, mild cognitive impairment, or subjective memory complaints) and

73 healthy volunteers completing the Cookie Theft picture description task. We show that eye tracking data is predictive of AD in a patient versus control classification task (AUC = .73). Furthermore, we show that using eye tracking data for this predictive task is complementary to using speech alone, as combining both modalities yields to the best classification performance (AUC=.80). Our results suggest that eye tracking is a useful modality for classification of AD, most promising when considered as an additional non-invasive modality to speech-based classification.

1. Introduction

Dementia is a progressive neurodegenerative process resulting in impaired cognition. It affects 47 million people worldwide, and it is one of the costliest diseases in developed countries (Arvanitakis et al., 2019; El-Hayek et al., 2019). Although dementia may be the end result of a number of processes, both alone or in combination, Alzheimer's disease (AD) is most common, contributing to approximately 60-80% of all cases.

The course of AD is insidious. Current evidence shows that pre-clinical pathological hallmarks of AD - amyloid plaques and neurofibrillary tangles - are present years before clinical symptoms occur (Vickers et al., 2016). Symptoms begin with subjective cognitive concerns, or subjective memory complaints (SMC) which may progress to mild cognitive impairment (MCI), where cognitive issues still do not impact day-to-day function. Dementia is defined by the point where cognitive impairment progresses to affect independent functioning, and progresses from mild to moderate to severe stages (Shaji et al., 2018). The most common presentations involve memory impairment, but additional domains, including language, visuospatial orientation, praxis (i.e. skilled movements) and personality changes may also occur. The median course of conversion between stages is highly variable and associated in part with a number of clinical, demographic and genetic factors. The four-year risk of conversion from SMC to MCI was 26% and SMC to dementia 14% in a larger meta-analysis of over 29000 individuals (Mitchell et al., 2014). There are no definitive disease-modifying treatments for AD. Clinical trials of potential disease-modifying therapies for AD have failed, in part, because neurodegeneration is too advanced at the time of diagnosis to change the course of decline due to the irreversible destruction of key brain pathways (Vickers et al., 2016; Sperling et al., 2014).

A successful disease-modifying drug for AD would be most likely to demonstrate an effect in individuals who do not yet have advanced neurodegenerative changes (Reiman et al., 2016; Sperling et al., 2014, 2011). Disease-modifying drug trials for dementia are increasingly focused on individuals with pre-clinical and very early stage disease, when successful therapeutics would be most likely to demonstrate an effect. However, current screening strategies for individuals for clinical trials who have pre-clinical disease, or are at higher risk for faster decline (and, conversely, exclusion of very low-risk individuals who are unlikely to progress), are highly inefficient and imprecise. Current estimates for pre-AD trial candidate screening costs are as high as $100,000 USD, while up to 80% of potential participants may be screen failures (Kolata, 2018). Thus, there is a strong need for efficient and accurate screening strategies to accelerate disease modifying-therapy development in AD.

A simple, accurate and non-invasive screening tool could help to accelerate curative treatments for AD by helping to select good trial candidates with early-stage or pre-clinical

disease and excluding healthy individuals, who will reduce the statistical power of a clinical trial. Machine learning models built on non-invasive patient data are obvious candidates to be used as non-invasive screening tools.

Previous research has shown the potential of using machine learning for the classification of AD/MCI versus healthy controls. Speech, in particular, has been extensively investigated for non-invasive AD risk-stratification tools (Pulido et al., 2020). To this end, eye movement data may serve as another means for risk stratification, as visuomotor impairments are present even in the prodromal phase of AD, before robust clinical symptoms develop. Recent work has studied the added predictive value of eye data in reading tasks alone (Biondi et al., 2017), or both language and eye data during reading tasks (Fraser et al., 2019), leading to promising results. However, no previous work has attempted classification of AD using eye movements collected during spontaneous speech tasks, nor the combination of eye tracking and language data in this context.

In this paper, we study if eye tracking data can be used for classification of AD/MCI versus controls on a spontaneous speech task. We begin with collecting data from memory clinic patients and healthy volunteers completing the "Cookie Theft" picture description task while their eye movements and speech are recorded. Using this dataset, we present two machine learning experiments. First, we investigate whether eye tracking data alone is predictive of patients versus controls. Second, we examine how classifiers based on eye tracking alone compare to language-based classifiers, and to classifiers using both eye tracking and language data. Our results show that eye tracking data alone can classify patients vs. controls in our dataset, compared to a standard majority class baseline. We also show that eye and language data are complementary, as combining eye and language features leads to the best classification performances (AUC =.80) compared to classifiers built on each modality alone. Furthermore, through an analysis of the most important features for our classifiers, we identify features with plausible clinical correlations with AD-related language and eye movement dysfunction.

Our main contributions are as follows: First, we build a dataset that includes eye tracking data in addition to speech data. Second, we are the first group to study eye tracking for classification of AD/MCI vs control using a spontaneous speech task, showing that unobtrusively collected eye tracking data is able to discriminate between patients and healthy controls in our dataset. Third, we are the first to show that eye tracking data is complementary to language data for this task.

Generalizable Insights about Machine Learning in the Context of Healthcare

1. **Need for contemporary, longitudinal and multimodal corpora.** Most related work is based on the Pitt corpus ("DementiaBank"), which is the largest publicly available speech corpus including individuals with AD and healthy controls. However, the dataset lacks longitudinal data, and clinical diagnostic practices have improved since the 1980s, when the corpus was collected in the field. We present a well-characterised contemporary dataset that includes eye gaze data in addition to speech data. The data collection is ongoing with scheduled 6-month follow-ups, which will constitute longitudinal data. We also report that our data collection methodology is well-tolerated in

the target population, which is essential for our goal of developing non-invasive risk stratification tools.

2. **Eye tracking as an additional data source for non-invasive risk stratification tools.** To date, language data has been the go-to approach for classification of dementia patients and healthy controls in spontaneous speech tasks, with very promising results. We show that eye tracking can be used as an additional, complementary, non-invasive modality for risk stratification.

3. **Classification models to provide insights for disease prediction.** Highly predictive features shed light on subtle dysfunction that may be difficult to clinically detect in the earlier stages of neurodegenerative disease and may aid in enrichment of clinical trial populations, or earlier diagnosis.

2. Related Work

2.1. Language Analysis and the DementiaBank Corpus

Previous groups have used ML algorithms and NLP techniques to develop automated classification for AD and/or MCI versus healthy controls with speech and language data. Many of traditional and deep learning models classifying AD have been developed using the Pitt DementiaBank corpus (Becker et al., 1994), the largest publicly available dataset incorporating speech transcripts of the Cookie Theft picture description task from 169 individuals with a clinical diagnosis of probable or possible AD, 19 with MCI and 99 healthy controls (aged 45-90). The corpus was collected between 1983 and 1988. Fraser et al. (2016) evaluated machine learning models incorporating acoustic and linguistic features to predict AD. With 370 custom lexical and acoustic features from the speech recordings, they achieved a predictive accuracy of 81.96% in distinguishing individuals with AD from healthy controls. Adding to this body of work, Field et al. (2017); Masrani (2018) proposed an approach in which they added a novel feature group based on the clinical observation that spatial neglect may be affected in individuals with AD (Drago et al., 2008). The authors divided the Cookie Theft picture into halves, strips and quadrants and computed features capturing the attention ratio between regions, examining the number of mentions of any given information unit (i.e., object or action) displayed within a region of the picture. Accuracy improved to 84.4%.

Other groups have further improved classification performance on the DementiaBank dataset using deep learning models. Kong et al. (2019) used a hierarchical attention recurrent neural network model and reported best performances when the model was trained on raw text data together with patient's age, leading to 86.9% classification accuracy. Karlekar et al. (2018), using single utterances from DementiaBank as individual data samples, found a 91% classification accuracy with part-of-speech-tagged utterances with a CNN-RNN model.

The Pitt corpus is limited in its lack of longitudinal data and thus is limited only to cross-sectional classification of AD or MCI versus control. A model incorporating lexical and acoustic information from the OPTIMA study, which included 15 individuals with autopsy-proven MCI or mild AD and 15 age- and education-matched healthy controls, found that

changes in connected speech at the MCI stage predicted conversion to AD, while semantic and lexical content, in addition to syntactic complexity, declined with disease progression (Ahmed et al., 2013).

2.2. Eye Movements

AD fundamentally alters ocular function (Molitor et al., 2015). The death of neurons, neurofibrillary tangles and amyloid plaques affect certain cell types in the neocortex leading to cortico-cortical disconnections. This primarily involves the temporoparietal association areas making AD patients prone to visual, attentional, and eye movement disturbances (Garbutt et al., 2008). Numerous studies have shown that AD patients exhibit abnormal pro-saccadic behavior, poor performance on antisaccade tasks, slowed pupillary responses, and impaired smooth pursuit (Molitor et al., 2015). In reading tasks, AD patients take longer to read a text, make more fixations, are more likely to re-read words and are much less likely to adaptively skip small and uninformative words (MacAskill and Anderson, 2016). The association between worsening visual task performance and disease progression is not well-characterized, although one study has suggested that antisaccade performance may predict AD severity (Crawford et al., 2005), and another that microsaccadic gaze intrusions are associated with worsened cognitive test performance in AD (Bylsma et al., 1995).

Recent work has shown the potential for eye movement data to be used in classification experiments of AD. Pavisic et al. (2017) reported a 95% accuracy using hidden Markov models that incorporated eye movement data from 36 individuals with young onset AD and 21 age-matched healthy controls. Participants performed three eye-tracking-specific tasks: fixation stability (stare at a static point for 10 seconds without blinking), pro-saccade (direct gaze towards a target as soon as it appears on screen), and smooth pursuit (follow a target while it moves on screen).

Other work has used eye movements collected during reading tasks. Biondi et al. (2017) collected data of 69 participants with probable AD (clinical impression and high-risk ApoE genotype) and 71 age-matched controls during a reading task, and reported 87.78% classification accuracy using a deep neural network model that incorporates information derived from fixations, saccades and sentence length from individuals who read high- and low-predictable sentences and proverbs. In another reading task-based study, Fraser et al. (2019) had 26 participants with MCI and 29 healthy volunteers read paragraphs from a standardized reading test silently and aloud while speech, including answers to comprehension questions, and eye movements were recorded, with additional speech-only data collected from the Cookie Theft Task. Best classification accuracy was 83% using a cascaded multimodal and multi-task classification approach incorporating custom lexical, acoustic, comprehension question-related, as well as eye tracking features related to saccades and fixations.

2.3. Addressing the Gap in the Literature

While previous studies have achieved reasonable classification accuracy, we contribute to the field by examining speech and simultaneous eye tracking during a picture description task (Cookie Theft) from a contemporary, prospectively collected and well-clinically-

characterized cohort. There are several advantages to our approach. First, we are investigating a multimodal approach to improve accuracy. Second, our task is aligned with that used in the Pitt corpus, which has been commonly leveraged to develop AD classification models based on speech data. Third, ours is a larger cohort (n=141) than other more contemporary prospective datasets examining speech or eye tracking data. Previous studies (Fraser et al., 2019; Pavisic et al., 2017; Toth et al., 2018; Biondi et al., 2017) are limited by small datasets, with the total number of participants in these studies ranging from 55 to 86. Additionally, our contemporary cohort is in alignment with current clinical practice for AD and MCI diagnosis, with a sample representative of current memory clinic populations and controls. Studies using the DementiaBank cohort (Fraser et al., 2016; Chen et al., 2019) achieved high accuracy but are limited by outdated dementia diagnostic criteria (Jack et al., 2011; Falk et al., 2018).

3. Data Collection

3.1. The Cookie Theft Picture Description Task

The Boston Cookie Theft picture description task is a well established speech task (Goodglass and Kaplan, 1972) (see Figure 1 in Section 4.2). It is a widely used and validated method for spontaneous speech assessment in a variety of clinical contexts, including Alzheimer's disease (Cummings, 2019). During the task, participants are shown the picture and are asked to describe everything they see in the scene using as much time as they would like.

3.2. Data Collection

Current cross-sectional data is presented, although recruitment and longitudinal follow-up is ongoing (target recruitment 250 clinic patients and 250 controls, with six-month longitudinal follow-up to 24 months). Patients were prospectively recruited from a specialized memory clinic with a catchment of 4 million and controls were recruited from the community, with efforts made for targeted recruitment for age and sex-matching with patients. All participants were fluent in English, able to provide informed consent and to carry on a spontaneous conversation, and aged 50 or older. Clinic patients had a diagnosis of subjective memory complaints (SMC), MCI or mild-moderate AD, without additional active psychiatric or neurological conditions. Diagnoses were made by expert clinicians with cognitive testing clinical data, and neuroimaging and laboratory data collected as per standard of care.

All participants gave informed consent. They completed the Montreal Cognitive Assessment test (MoCA), a guideline-recommended ten-minute pen-and-paper cognitive screening test for assisting health professionals in the diagnosis of MCI and AD (Nasreddine et al., 2005; Gauthier et al., 2012; Cordell et al., 2013), and a demographic and medical history questionnaire, with responses cross-checked against the clinic electronic medical record (full results are reported in the Appendix A). During the Cookie Theft task, participants were seated at the testing platform consisting of a computer monitor with a Logitech C922x ProStream video/sound recorder and an infrared eye-tracker (Tobii-Pro X3-120) affixed at the bottom of the monitor to record gaze and pupil size data. Participants then performed

a standard 9-point eye-tracker calibration, as well as a pupil baseline collection procedure which consisted in having the participant relax and fixate on a blank screen for 10 seconds. Following successful calibration, participants performed the Cookie Theft picture description task. As we used a non-intrusive, remote eye-tracking device that does not restrict participants movements, they were asked to keep looking at the screen while they performed the task, and avoid looking at the experimenter.

Following the assessment, participants were asked to complete a questionnaire to rate their experience with the assessment. Specifically, participants were asked about the ease of use, acceptability, and attitude towards speech and eye tracking gaze technology on a 4-point Likert scale, to determine the usefulness and scalability of the technology for routine assessment. This was introduced midway into the study. Overall, patients' experience with the assessment was perceived positively. Out of the 56 patients that completed the questionnaire, 93% felt comfortable and 91% felt relaxed during the assessment. Additionally, 96% of them reported to be willing to repeat the assessment again in the future, with a total of 93% participants willing to repeat the assessment on a monthly basis. Refer to the Appendix B for the full responses on the experience with technology questionnaire.

3.3. Cohort Characteristics

Sixty eight memory clinic patients (34 with AD, 22 with MCI, 7 SMC, and 5 with mixed dementia) and 73 healthy volunteers were recruited from May 2019 to March 2020. Their demographics, clinician diagnosis, and MoCA scores are shown in Table 1.

Table 1: Demographic and clinical data. Additional data on cohort characteristics can be found in the Appendix A

		Patient	Control
Total number	N	68	73
Participant sex	Male	34	22
	Female	34	51
Age at enrollment	Average	71.6	64.9
	Range	52-96	50-83
	Standard deviation	9.26	9.93
Expert clinician diagnosis	AD	34	0
	MCI	22	0
	SMC	7	0
	Mixed dementia†	5	0
MoCA score (0 - 30 scale)	Available scores	n=59	n=71
	Average	20.25	27.15
	Range	5-29	19-30
	Standard deviation	5.44	2.73

†characteristics of AD and vascular dementia

For this analysis, participants were grouped into "patient" or "control" categories. Participants were coded as "patient" if they were a patient recruited from the memory clinic, or "healthy" if they were recruited as the patient's companion or from the community. We chose to group these participants as this study aims to identify highly predictive speech and eye movement features shared across each stage of the disease based on clinical characterizations. Features characteristic of manifest AD that are also found in people with MCI or SMC may predict with higher likelihood conversion to AD, and may be helpful for future risk-stratification work.

3.4. Data Pre-Processing

Following data collection, eye and speech recordings underwent pre-processing procedures. Speech data was either transcribed manually (111/141), or with Google Cloud speech-to-text service[1] (31/141), the latter being introduced later into the study due to institutional contract and ethics approvals. Following automatic transcription, transcripts were manually verified for accuracy by human transcribers. Eye tracking data was exported using the Tobii Pro Studio software[2], comprising fixations (points of gaze on the screen), saccades (quick movements between fixations), and pupil size. Pupil size data was baseline-adjusted (Iqbal et al., 2005) by subtracting the mean pupil size collected during the pupil baseline calibration procedure described in Section 3.2.

4. Classification of Patients and Controls Using Eye Tracking Data

As mentioned in the introduction, one of the contributions of this paper is that we are the first to explore the potential value of leveraging eye tracking data to detect dementia from the Cookie Theft picture description task. As a first step in our analysis, we verify whether just leveraging eye movements captured during the image description task can classify patients (SMC/MCI/AD) versus healthy controls. To this end, we compare the performance of two different feature sets, one that is task-agnostic, and one that relies on identifying relevant areas of interest (AOI) on the Cookie Theft picture (task-specific). We also compare the performance of these feature sets using a larger but more noisy dataset compared to a smaller but cleaner dataset, in order to study the data quality-quantity tradeoff. Section 4.1 describes the two feature sets we use for the analysis. Section 4.2 explains the experiment settings. Section 4.3 discusses the results.

4.1. Eye tracking features

In order to capture users' eye-movement and pupil behaviour, we compute a set of summary statistics on the fixation, saccades and pupil size data, following the standard approach in related work (Toker et al., 2017, 2019; D'Mello et al., 2012; Lallé et al., 2016; Martínez-Gómez and Aizawa, 2014). We processed the eye tracking data using the Eye Movement Data Analysis Toolkit (EMDAT[3]), an open source library written in Python.

1. https://cloud.google.com/speech-to-text
2. https://www.tobiipro.com/product-listing/tobii-pro-studio/
3. https://www.cs.ubc.ca/~skardan/EMDAT

EMDAT produces a comprehensive set of eye tracking metrics specified over the entire display (task-agnostic), and over specific Areas of Interests (AOIs; task-specific). We defined the following 13 AOIs to encode elements in the Cookie Theft picture, which are shown in Figure 1: cookie, cookie jar, boy, girl, woman, stool, plate, dishcloth, water, window, curtain, dishes, sink. Note that the AOIs defined for the task-specific features are analogous to the "information units" Croisile et al. (1996) used in language analyses on the Cookie Theft picture description task. The complete list of metrics grouped in the task-specific and task-agnostic feature sets are described in Table 2.

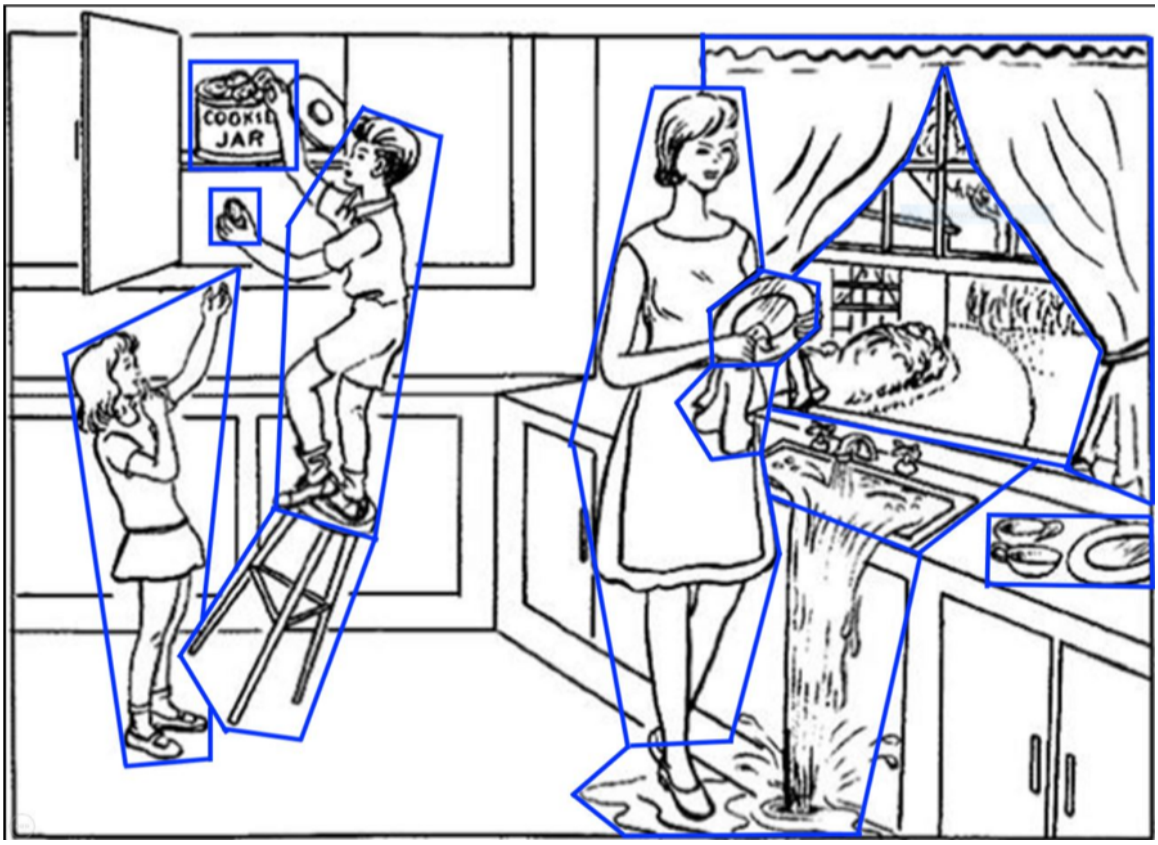

Figure 1: Areas of interest (AOIs) in blue, defined over the Cookie Theft picture.

4.2. Experiment settings

Datasets: For 120 out of the 141 participants in our cohort, the calibration of the eye-tracking device was exceptionally good (the nine points were calibrated successfully), whereas for 12 of these participants, at least one of the calibration points was not successfully calibrated. To estimate the impact of data quality-quantity tradeoff, we evaluate the classifiers on the full dataset of 141 participants, called the Full dataset, as well as the subset of 120 participants for which we have cleanest eye tracking data quality, called the Clean dataset.

Table 2: Features computed from the eye tracking data.

Feature set	Signal component	Metric
Task-agnostic (E_TA)	fixation(6)	count, rate duration: sum, avg, stdev, max
	saccade(22)	count, rate distance: sum, avg, stdev, max duration: sum, avg, stdev, max speed: avg, stdev, min, max absolute angle: sum, avg, stddev, rate relative angle: sum, avg, stddev, rate
	pupil size (6)	avg, stdev, max, min, start, end
Task-specific (E_TS)	fixations on AOIs (9)	count, rate, proportion duration: sum, avg, stdev, max time to first fixation, time to last fixation
	transitions to AOIs (2)	count, proportion
	pupil size in AOI (6)	avg, stdev, max, min, start, end

Classifiers: We test three different classification algorithms, Logistic Regression (LR), Random Forest (RF), and Gaussian Naive Bayes (GNB), which reported top performances in most closely related work Masrani (2018). We use scikit-learn, a python package for machine learning, to perform classification.

Baseline: To see whether eye tracking features are sufficiently accurate classifiers on their own, we use a zero rule classifier (B), which always predicts the majority class in training data, as a baseline for comparison.

Feature sets: We compare the classification performance of classifiers using task-agnostic (E_TA) features only, task-specific features (E_TS) only, and the combination of both (E).

Evaluation: We evaluate classifiers using a stratified 10-fold cross-validation approach, which is repeated 10 times (runs) on different stratified splits to strengthen the stability and reproducibility of the results. We report classification performance in terms of area under the ROC curve (AUC), averaged over the 10 folds and the 10 runs. At each fold of cross-validation, we perform correlation feature selection Hall (1999) to remove highly pairwise correlated features (Pearson r $>.85$) as well as features with a very low correlation with the outcome (Pearson r $<.2$).

4.3. Results

Table 3 reports the performance of classifiers using task agnostic (E_TA) features alone, task-specific (E_TS) features alone, as well as the combination of both feature sets (E). To statistically compare the classification results we run a 3-Way ANOVA with AUC as the dependent variable and Dataset, Feature set, and Classifier as factors.

Classification performance of eye tracking features for classification. We find a main effect of classifier on AUC scores ($F(2,231) = 147.33$, $p<.000$, $\eta_p^2=.56$). Pairwise

Table 3: Performance results in terms of AUC (± sd) from the classification experiments using the eye tracking features. Bold entries indicate highest classification performances for each feature set. B: Zero-rule baseline, GNB: Gaussian Naive Bayes, LR: Logistic Regression, RF: Random Forest, E_TA: Task agnostic features , E_TS: Task-specific features, E: Task agnostic features + Task-specific features.

		Classifier			
Feature set	Dataset	B	GNB	LR	RF
E_TA	Clean	.50 ± .00	.57 ± .05	.53 ± .04	.52 ± .04
	Full	.50 ± .00	**.62** ± .05	.59 ± .05	.54 ± .02
E_TS	Clean	.50 ± .00	**.70** ± .02	.70 ± .03	.70 ± .05
	Full	.50 ± .00	.67 ± .02	.67 ± .03	.68 ± .02
E	Clean	.50 ± .00	.70 ± .02	.70 ± .03	**.73** ± .04
	Full	.50 ± .00	.66 ± .02	.67 ± .02	.70 ± .02

contrast comparisons show a significantly[4] higher AUC score for each of the classifiers (GNB, LR, RF) compared to the baseline, with no statistically significant differences among them. This result indicates that all three classifiers are equally good when classifying patients vs. healthy controls from eye tracking data.

Task-agnostic vs. task-specific features. We find a main effect of Feature set on AUC ($F(3,231) = 202.43$, $p<.0001$, $\eta_p^2=.64$). Pairwise contrast comparisons indicate significantly better performance of both task-specific features (E_TS) and combined features (E) as compared to task-agnostic features (E_TA), while no statistical difference was found between E_TS and E features. This result supports that encoding task-specific eye movements leads to the best classification performance in the Cookie Theft picture description task, with no added value provided by task-agnostic features.

Impact of data quality-quantity tradeoff. Last, we find a significant interaction effect between Feature set and Dataset on AUC ($F(2,231) = 10.10$ $p<.0001$, $\eta_p^2=.08$). Pairwise contrast comparisons indicate that for E_TA features, using the Full dataset (i.e., more data) increases classification performance as compared to the Clean dataset. In contrast, for E_TS and E feature sets, the Clean dataset (i.e., less data but better quality) leads to statistically better performances compared to the Full dataset. This result highlights the importance of eye tracking data quality when considering task-specific features, which makes sense as these features rely on accurately tracking the patient's eyes over specific regions on the image. In contrast, it is important to gather more data, even if noisier when using task-agnostic features alone, because they are less reliant on specific regions of the screen, but encode more generic eye-movement information, not directly related to the task at hand.

Given the results reported in this Section, we select the Feature set and Dataset that led to highest classification performance (E feature set using the Clean dataset) to move on to

4. in this paper significance is reported at $p < 0.05$ after applying Benjamini and Hochberg procedure to adjust for the false discovery rate Benjamini and Hochberg (1995)

the following analysis, which involves comparing and combining eye tracking and language data for classification.

5. Combination and Comparison of Eye Tracking Features and Language Features

As eye tracking features show effectiveness for classifying patients from healthy controls in Section 4, in this Section we investigate whether eye tracking features will help classification when combined with language-based features, which have already been proven to be effective. Specifically we evaluate multimodal classifiers that leverage eye tracking and language features using two kinds of fusion schemes, the early fusion and late fusion. Section 5.1 introduces language features. Note that for eye tracking features, we use the E feature set defined in Section 4.2. Section 5.2 explains the methods we use for early fusion and late fusion. Section 5.3 explains the experimental settings, and Section 5.4 discusses the results.

5.1. Language Features

For language features, we use a comprehensive set of language features from previous work, most notably from Fraser et al. (2016). This set contains two different sub-groups: text and audio. The text features include part-of-speech (15), context-free-grammar rules (44), syntactic complexity (27), vocabulary richness (4), psychologistic (5), repetitiveness (5), and information units (40). The audio features include 172 acoustic features as in Masrani (2018).

5.2. Fusion Models

To combine data from different modalities for our study, we explore both early fusion and late fusion methods. These are the two kinds of fusion schemes generally used for multimodal approaches.

For early fusion, we concatenate features from the two modes, and make a single feature vector to learn a classifier. This approach is simple and it allows modeling interactions among features depending on the classifier. Our early fusion model is shown in Figure 2.

The late fusion scheme combines predictions from each modality at the decision level (see Figure 3). We use a widely established late fusion approach named "voting by averaging", in which predictions are made by averaging the outputs of different learning algorithms (with heterogeneous model representations) to a single dataset (Battiti and Colla, 1994). In our case, we use a slightly modified approach, in which we average the prediction probabilities produced by a single learning algorithm, but applied to different data modalities, as in Fraser et al. (2019).

5.3. Experiment Settings

In this experiment, we aim to investigate whether eye tracking features in combination with language features outperform eye tracking features alone, as well as language features alone.

Dataset: We evaluate the multimodal models using the Clean dataset, as it was the one for which we achieved the overall highest classification performance when using eye tracking features only in Section 4.

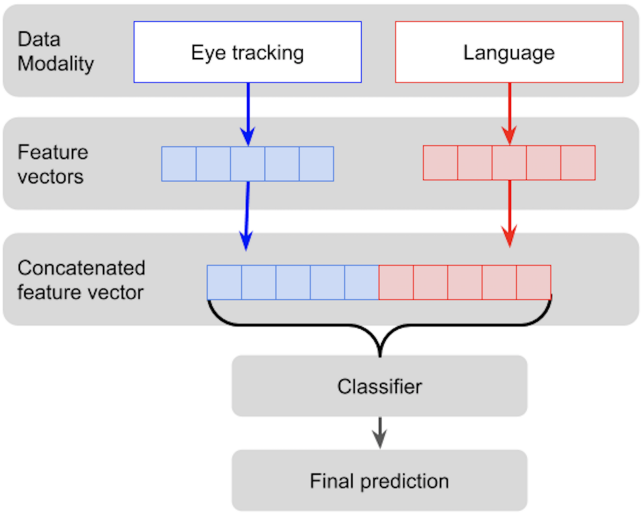

Figure 2: In the early fusion, features of the two modes are concatenated into a single feature vector to learn a classifier.

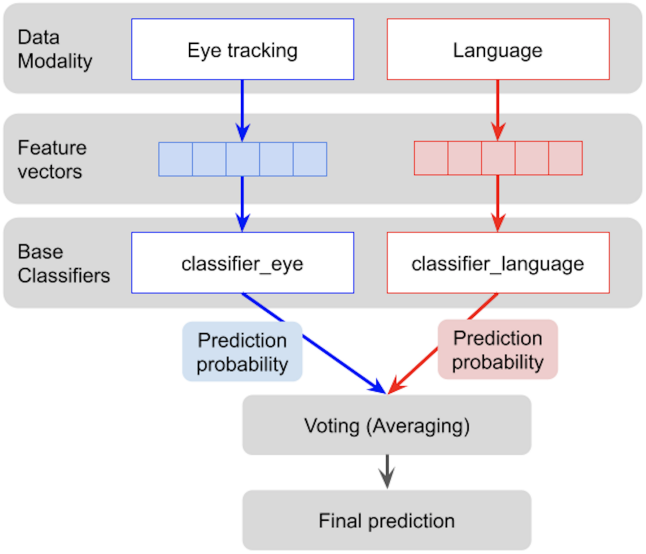

Figure 3: In the late fusion, predictions from individual classifications are averaged at the decision level

Classifiers: We use the same classifiers as in the previous experiment, that is, Logistic Regression (LR), Random Forest (RF), and Gaussian Naive Bayes (GNB). In this experiment, we drop the majority class baseline as we have shown that our eye-tracking-based models statistically outperform that baseline.

Feature sets : To see the effects of multimodal models, we compare both early fusion (Early_E+L) and late fusion (Late_E+L) models against models that use features from a single mode, i.e., models using eye tracking features only (E) and language features only (L). Note that L corresponds to the feature set described in Section 5.2, which is a state-of-the-art language feature set using traditional algorithms on the DementiaBank dataset, and E corresponds to the model with highest classification performance in Section 4.

Demographics Baseline: In addition to the above, we also compare the multimodal models to a baseline built on participants' demographics features. Demographic features (particularly age, sex, education, and medical history) can be obtained non-invasively, and have shown to be predictive of dementia in previous studies (Calvin et al., 2019; Kim et al., 2017), providing a more realistic baseline than the majority class baseline used in Section 4. For the demographics baseline we use standard demographics characteristics, such as sex and age (see Table 1), as well as participants' medical history (e.g., history of concussion, stroke etc.), level of education, and alcohol and tobacco consumption. Details and descriptive statistics on the features included in the demographics baseline are reported in the Appendix A (items marked with an asterisk).

Evaluation: As in Section 4, we evaluate the models using stratified 10-fold cross-validation repeated ten times, and report classification performance in terms of AUC. We perform the same correlation feature selection procedure as described in Section 4.

5.4. Experiment Results

Table 4 Summarises the results of the multimodal experiments. To statistically compare the performance of the early and late fusion approaches to the performance of the single modality models and to the demographics baseline, we run a 2-way ANOVA on the AUCs with Classifier and Feature set as factors.

Table 4: Performance results in terms of AUC (± sd) of the classification experiments using the multimodal models. Bold entries indicate highest classification performances for each feature set. GNB: Gaussian Naive Bayes, LR: Logistic Regression, RF: Random Forest, D: Demographics baseline, E: Eye tracking features, L: Language features, Early_E+L: Early fusion of E and L, Late_E+L: Late fusion of E and L

	Classifier		
Feature set	GNB	LR	RF
D	**.66** ± .01	.65 ± .02	.62 ± .03
E	.70 ± .02	.70 ± .03	**.73** ± .04
L	**.77** ± .02	.72 ± .03	.71 ± .02
Early_E+L	**.77** ± .01	.73 ± .03	.71 ± .04
Late_E+L	**.80**± .02	.77 ± .03	.75 ± .04

The ANOVA result shows a main effect of Feature Set on AUC ($F(4,135)=90.67$, $p<.0001$, $\eta_p^2=.73$). Pairwise contrast comparisons indicate that the models using eye tracking features (E), language features (L), and a combination of those (Early_E+L and Late_E+L) all

outperform the demographics baseline. This indicates that eye movements captured during the Cookie Theft picture description task are more predictive than simple demographics and medical history characteristics alone, and so are their combination with language data.

Furthermore, pairwise contrast comparisons indicate that Late_E+L significantly outperforms the language data alone L, which in turn outperforms the model using eye tracking data alone E. Early_E+L, on the contrary, does not statistically outperform the state of the art L model. These results indicate that 1) eye and language modalities are complementary, as combining eye and language features lead to the best classification performances; 2) this is only achieved using late fusion methodology, while simply concatenating the features (Early_E+L) does not outperform the language-only (L). Possible reasons for late fusion's better synergy between different modalities are discussed in Section 6. We also tried adding the demographics features D to the multimodal models Early_E+L and Late_E+L, but this did not lead to statistically improved classification performance.

In addition, the ANOVA result shows a main effect of classifier on AUC ($F(2,135)=25.43$, $p<.0001$, $\eta_p^2=.57$). Pairwise contrast comparisons indicate that overall, GNB statistically outperformed both LR and RF, while no statistical difference was found between LR and RF. One possible explanation is that GNB has been shown to perform comparably well for relatively smaller datasets, compared to other algorithms such as LR and RF (Mitchell, 1997).

6. Further Insights on the Classification Experiments

In order to provide further insights from our classification experiments, we look into which are the most predictive features used by the best performing classifier (GNB). GNB classifiers estimate $P(class \mid data)$ by learning conditional probabilities of each feature given a class, $P(feature \mid class)$, under the conditional independence assumption of features. As a result, we can obtain $P(feature \mid class)$ from a trained model using GNB (Mitchell, 1997). We assume that highly discriminative classification features have separate distributions between the classes (patient and control). A feature is considered to be discriminative when the difference in means is large, relative to the variance of the distributions. As a way to estimate discrimination power, we use the following measure, which is also used in Linear Discriminant Analysis (LDA) (Belhumeur et al., 1997):

$$d = \left| \frac{\mu_{HC} - \mu_P}{\sqrt{\sigma_{HC}^2 + \sigma_P^2}} \right|$$

, where μ_{HC} is the mean for healthy controls, μ_P is the mean for patients, σ_{HC}^2 is the variance for healthy controls, and σ_P^2 is the variance for patients.

We train GNB on the whole dataset using the features in the early fusion model (Early_E+L), as well as the features from each individual mode (E and L) used in the late fusion model (Late_E+L). Table 5 shows the top 10 ranked features for each model. For each feature on the table, we report whether the mean values in the patient group are higher (+) or smaller (-) than in the control group. Note that some language features are not easy to interpret (e.g., frequency-related audio features). Thus, for simplicity, we only

specify language features related to information units (see Section 5), which are analogous to the eye tracking AOI features (see Section 4).

Table 5: Top 10 most important features for Eye and Language, as well as for concatenation of these modalities. Features with higher mean values in the patient group and control group are labeled as (+) and (-) respectively.

Feature Set	Top 10 important features
E	number of transitions to other AOIs: dish to window (+), window to girl (+), window to plate (+), jar to curtain (+), boy to cookie (-), window to window (-), sink to girl (-), cookie to girl (-), proportion of fixations on AOI: cookie jar (+) saccade absolute angle: stdev (+)
L	8 acoustic features, 2 text features, 0 information unit related features
Early_E+L	8 acoustic features, 2 information unit features: number of mentions to water overflowing (-), mentioning girl at least once (-)

Classification performance of early vs. late fusion In Section 5.3 we reported that combining eye tracking (E) and language (L) features led to improved performance only when using late fusion (Late_E+L), while simply concatenating the features (Early_E+L) did not outperform the language-only (L) model. By examining the top 10 ranked features for Early_E+L (Table 5-bottom), we observe that there are only features from the language modality. This reveals that in this case early fusion overweighs language as the strongest modality. On the contrary, late fusion by combining the classification outputs from individual modes at the decision level, is able to better integrate the modalities in a complementary way.

Observations for clinical insights from eye tracking data. To investigate how the eye tracking model (E) performs, we turn our attention toward eye tracking features (Table 5-top). The best eye tracking features are related to information units (elements specific to the Cookie Theft image), confirming the importance of task-specific features for prediction discussed in Section 4.3. In particular, highly predictive features characteristic of patients involve transitions between different AOIs representing information units of the picture (dish to window, window to girl, jar to curtain), as if patients are jumping back and forth between different elements. Patients also have a higher proportion of visits to the cookie jar. In contrast, controls show more transitions within the window. Healthy controls also have more transitions to the girl (sink to girl, and cookie to girl), which is further supplemented by more mentions of the girl by controls (Table 5-bottom) .

Note that there were no significant differences between the time patients (Mean = 69.83s) and healthy controls (Mean = 82.01s) spent completing the task (t (107.2) = -1.26, p= .056). Interestingly however, most top eye tracking features are related to the total count of transitions between information units, within the span of the task. This means that while

both groups look at the image for the same duration, they process the information in the image in different ways.

Observations for clinical insights from eye tracking and language data. To gain insight about the differences in information-processing strategies between patients and healthy controls, we look at the top-ranked features for models E (Table 5-top) and E+L (Table 5-bottom) since they both include information unit-related features. We found that healthy controls speak more references to information units than patients, despite both groups spending similar amounts of time on the task. Healthy controls also described the water overflowing more than patients, even though none of the top eye-movement related features involved water. Both healthy controls and patients may be observing the water and the rest of the image in a similar way, with only the controls actually mentioning information units or describing the water.

The discussion of the clinical implications from the above observations is provided in the following section.

7. Discussion

With the goal of developing an accurate, non-invasive screening tool for AD clinical trials, we presented an analysis of eye movements and speech recordings on a prospectively well-phenotyped multimodal corpus based on the established Cookie Theft picture description task. A limitation of similar cohorts (i.e. DementiaBank) is that the participant data phenotyping is limited, providing only diagnostic codes assigned to each sample. Collecting more comprehensive medical history, cognitive testing results, demographic and imaging data allows our cohort to be even more useful for investigating classification of pre-clinical dementia, MCI and AD and is helpful in determining whether the cohort derived to train the algorithm is reflective of the clinical target population.

For the first time, we analyzed multimodal data captured during the Cookie Theft picture description task. Multimodal data (i.e. language and eye tracking) is promising in dementia as its pathology involves the decline of multiple different cognitive domains. Capturing multiple domains increases the likelihood of detecting an at-risk individual. We found that this methodology of data collection is feasible and well-tolerated in the target population, with a large majority of participants feeling comfortable and relaxed during the assessment. Furthermore, as risk-stratification tools would likely involve multiple assessments longitudinally, it is reassuring that a majority of memory clinic patients were willing to repeat the assessment again in the future, as often as on a monthly basis.

Our experimental results are promising for the goal of developing a non-invasive risk stratification tool, as we demonstrate that an additional novel non-invasive data modality (eye tracking) can be used to classify patients in our cohort. Specifically we show that models using eye tracking features alone were discriminative in our dataset, reaching a peak performance of AUC = .73. Eye-tracking technology is becoming more cost effective and accessible (Tobii, 2017), suggesting that reliable eye tracking data may be ubiquitously available in the coming years. To this end, we investigated data quality-quantity tradeoff, and its impact on the predictive performance of eye tracking data. As the most predictive features in this task were related to task-specific elements in the picture, and given the rel-

atively small dataset size, our analysis suggests that under these circumstances, prioritizing data quality over quantity is more useful for accurate classification.

Most importantly, we showed that this modality is complementary to language, which is the most well studied modality for this classification task. In fact, the highest performance from language features on our dataset was improved from .77 to .80 AUC when including eye-movement-related information. These are encouraging results as, beyond the specifics of the increment in performance, it demonstrates that collecting simultaneously different modalities can lead to improved classification of AD/MCI/SMC vs. healthy controls. Further it uses data collected from non-invasive (in comparison to blood or cerebral spinal fluid-derived markers), modalities that do not make use of advanced neuroimaging (such as resource-intensive positron emission tomography (PET), single photon-emission CT (SPECT) or MRI), do not require expert interpretation in clinical context (such as laboratory biomarkers or neuroimaging), and pose essentially no risk to patients.

As we discussed in Section 6, in our study the majority of highly predictive gaze features were related to transitions between information units, while time spent on the task remained similar between groups. A higher proportion of transitions in the AD group may be attributed to impaired visuospatial processing and poorer short-term or working memory in the patient group, or poorer executive function. Patients may be transitioning more to information units, as they are not recalling what they had just looked at; in part they may also have a less organized approach to assessing the picture. This correlates with previous studies finding that individuals with AD have impaired visuospatial short-term memory, detectable in delayed reproduction tasks (Liang et al., 2016) and visual paired comparison tasks (Zola et al., 2013). While these tasks both specifically test for memory, our features may be inadvertently detecting impairments in visuospatial memory. "Misbinding" of features between items (eg. mixing up characteristics between items) is proposed as the mechanism for impair working visual and verbal memory in AD (Zokaei and Husain, 2019). Early deficits in executive dysfunction (Guarino et al., 2018) have also been increasingly well-described in AD.

Further, when comparing the top language features in patients and controls, controls spoke more about information units overall despite both groups looking at the picture for a similar duration. Alzheimer's patients commonly develop logopenic-type progressive aphasia: gradual language impairments characterized by word-finding difficulty, word-retrieval pauses, and loss of fluency (Mesulam et al., 2014). Eye movement patterns similar to healthy controls with discrepancies in information unit mentions may be attributed to progressive dysphasia and word-finding difficulties. This insight would not be possible in a single-mode assessment, and warrants future investigation.

7.1. Limitations

This paper provides proof-of-concept that eye tracking features are effective, independently discriminative, and complementary to well-researched language features. We used existing speech analysis algorithms in our study as the goal of solely achieving better performance, while desirable, was secondary to the aim of evaluating the utility of eye tracking. For this reason, and given our limited dataset, we did not pursue fine tuning hyper-parameters nor did we explore more advanced feature selection.

Weighing the trade-off between data quality and quantity for eye tracking features in Section 4, we discussed that for our dataset, data quality seems to be more important than quantity. However, while this could be informative for future research, we need to be cautious in generalizing these findings, as both the "Full" and "Clean" datasets are relatively not that large. With a dataset orders of magnitude larger, a high quantity of data could overcome poorer quality data. As our data collection is ongoing, and the dataset grows in size, we will be studying how this trade-off changes.

The technology we used for eye tracking is not sufficiently sensitive to characterize differences in microsaccadic eye movements, which may also be discriminative between patients and healthy controls. However, to capture these high resolution eye movements, head movements need to be restricted using a chin rest (e.g., EyeLink 1000[5]), which is uncomfortable and poorly tolerated by elderly subjects. Our approach is a more feasible eye tracking data collection procedure for this population, and our results show that this approach is predictive, and at the same time is accepted by the target population.

While our memory clinic patient cohort is well-characterized - with expert diagnosis, test scores, clinical neuroimaging, and laboratory data - our healthy control cohort currently lacks a similar validation, making it possible that some "healthy" controls may in fact have undiagnosed neurodegenerative issues. Incorporating neuroimaging (3D volumetric MRI) and polygenic hazard risk-stratification(Tan et al., 2017) for healthy controls should increase our confidence that our controls are in fact healthy. Furthermore, despite clinician diagnosis being based on the most updated neuroimaging, biochemical, clinical, and cognitive testing criteria, they can still be incorrect. In fact, a large post-mortem histopathological study found that AD is often misdiagnosed, with diagnostic sensitivity ranging from 70.9% to 87.3%, and diagnostic specificity ranging from 44.3% to 70.8% (Beach et al., 2012), introducing some uncertainty into our analysis. This limitation is not specific to our investigation alone, as many other dementia trials and studies rely primarily on expert clinical diagnosis. As diagnostic tools and technology evolve, the diagnostic accuracy of patients in our cohort should only increase. We are also continuing targeted recruitment to ensure optimal age- and sex-matching between controls and patients.

Eye movement, pupillary and speech changes can occur with a number of neurological conditions. Though the characteristics of these changes tend to differ clinically between dementia types (i.e., AD versus lewy body dementia, frontotemporal dementia, etc., which are due to different pathologies and would require different disease-modifying therapies), there may be some overlap with regards to changes in speech and eye-movement/pupil changes. Future work with recruitment of patients with these diagnoses is required to determine if our platform may discriminate between these different pathologies.

Finally, although our current methods report lower classification performances than in some other studies distinguishing AD/MCI from healthy controls, they have not been validated in larger, contemporary and well-characterized corpora and their results may in part be to particular characteristics of their training sets.

5. https://www.sr-research.com/eyelink-1000-plus/

7.2. Future Work

Recruitment for this study is still ongoing, and the study cohort is increasing in size. Participants have also agreed to complete follow-up assessments every 6-months over two years, to track disease progression, speech, and language changes over time. Reassessing individuals to identify those with progressive cognitive decline (shift of at least one category from control>SMC>MCI>mild AD>moderate AD) will allow us to identify features predictive of cognitive decline, with the goal of risk-stratifying individuals.

Future assessments with our cohort will include comparing the discriminative ability of our approach to distinguish disease stages (e.g., SMC versus MCI versus AD), and, once genetic and MRI data is fully collected from our healthy controls, assessing classification accuracy for high-risk pre-clinical (totally asymptomatic) individuals currently characterized as healthy controls. We are currently also welcoming controls with a lower age inclusion criteria ($>$ 19 years old), which will be helpful in that future work will also enable comparisons with young pre-clinical individuals carrying high-risk genes for young-onset AD.

Lastly, as our cohort size increases, we will focus more on technical improvements of the machine learning models to more effectively leverage eye tracking and language features, including new multimodal feature design, as well as representation learning algorithms and neural network approaches, in order to improve classification performance.

References

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

Zoe Arvanitakis, Raj C Shah, and David A Bennett. Diagnosis and management of dementia: Review. *JAMA*, 322(16):1589–1599, October 2019.

Roberto Battiti and Anna Maria Colla. Democracy in neural nets: Voting schemes for classification. *Neural Networks*, 7(4):691–707, January 1994. ISSN 0893-6080. doi: 10.1016/0893-6080(94)90046-9. URL http://www.sciencedirect.com/science/article/pii/0893608094900469.

Thomas G Beach, Sarah E Monsell, Leslie E Phillips, and Walter Kukull. Accuracy of the Clinical Diagnosis of Alzheimer Disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010. *J. Neuropathol. Exp. Neurol.*, 71(4):266–273, April 2012.

J T Becker, F Boller, O L Lopez, J Saxton, and K L McGonigle. The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis. *Arch. Neurol.*, 51(6):585–594, June 1994.

Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Juan Biondi, Gerardo Fernandez, Silvia Castro, and Osvaldo Agamennoni. Eye-Movement behavior identification for AD diagnosis. *arXiv:1702. 00837 [cs, q-bio]*, February 2017.

F W Bylsma, D X Rasmusson, G W Rebok, P M Keyl, L Tune, and J Brandt. Changes in visual fixation and saccadic eye movements in Alzheimer's disease. *Int. J. Psychophysiol.*, 19(1):33–40, February 1995.

Catherine M Calvin, Tim Wilkinson, John M Starr, Cathie Sudlow, Saskia P Hagenaars, Sarah E Harris, Christian Schnier, Gail Davies, Chloe Fawns-Ritchie, Catharine R Gale, John Gallacher, and Ian J Deary. Predicting incident dementia 3-8 years after brief cognitive tests in the UK biobank prospective study of 500,000 people. *Alzheimers. Dement.*, 15(12):1546–1557, dec 2019.

Jun Chen, Ji Zhu, and Jieping Ye. An Attention-Based hybrid network for automatic detection of alzheimer's disease from narrative speech. In *Interspeech 2019*, pages 4085–4089, ISCA, September 2019. ISCA.

Cyndy B Cordell, Soo Borson, Malaz Boustani, Joshua Chodosh, David Reuben, Joe Verghese, William Thies, Leslie B Fried, and Medicare Detection of Cognitive Impairment Workgroup. Alzheimer's association recommendations for operationalizing the detection of cognitive impairment during the medicare annual wellness visit in a primary care setting, 2013.

Trevor J Crawford, Steve Higham, Ted Renvoize, Julie Patel, Mark Dale, Anur Suriya, and Sue Tetley. Inhibitory control of saccadic eye movements and cognitive impairment in alzheimer's disease. *Biol. Psychiatry*, 57(9):1052–1060, May 2005.

Bernard Croisile, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and language*, 53(1):1–19, 1996.

Louise Cummings. Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia. *Pragmatics and Society*, 10:151–174, March 2019.

Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, 70(5):377–398, 2012.

Valeria Drago, Paul S Foster, Raffaele Ferri, Debora Arico, Bartolo Lanuzza, and Kenneth M Heilman. Distractibility and alzheimer disease: The "neglected" phenomenon, 2008.

Youssef H El-Hayek, Ryan E Wiley, Charles P Khoury, Ritesh P Daya, Clive Ballard, Alison R Evans, Michael Karran, José Luis Molinuevo, Matthew Norton, and Alireza Atri. Tip of the iceberg: Assessing the global socioeconomic costs of alzheimer's disease and related dementias and strategic implications for stakeholders. *J. Alzheimers. Dis.*, 70(2):323–341, 2019.

Nathan Falk, Ariel Cole, and T Jason Meredith. Evaluation of Suspected Dementia. *Am. Fam. Physician*, 97(6):398–405, March 2018.

Thalia S Field, Vaden Masrani, Gabriel Murray, and Giuseppe Carenini. IMPROVING DIAGNOSTIC ACCURACY OF ALZHEIMER'S DISEASE FROM SPEECH ANALYSIS USING MARKERS OF HEMISPATIAL NEGLECT. *Alzheimers. Dement.*, 13(7): P157–P158, July 2017.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *J. Alzheimers. Dis.*, 49(2):407–422, 2016.

Kathleen C Fraser, Kristina Lundholm Fors, Marie Eckerström, Fredrik Öhman, and Dimitrios Kokkinakis. Predicting MCI status from multimodal language data using cascaded classifiers. *Front. Aging Neurosci.*, 11:205, August 2019.

Siobhan Garbutt, Alisa Matlin, Joanna Hellmuth, Ana K Schenk, Julene K Johnson, Howard Rosen, David Dean, Joel Kramer, John Neuhaus, Bruce L Miller, Stephen G Lisberger, and Adam L Boxer. Oculomotor function in frontotemporal lobar degeneration, related disorders and Alzheimer's disease. *Brain*, 131(Pt 5):1268–1281, May 2008.

Serge Gauthier, Christopher Patterson, Howard Chertkow, Michael Gordon, Nathan Herrmann, Kenneth Rockwood, Pedro Rosa-Neto, and Jean-Paul Soucy. Recommendations of the 4th canadian consensus conference on the diagnosis and treatment of dementia (CCCDTD4). *Can. Geriatr. J.*, 15(4):120–126, December 2012.

Harold Goodglass and Edith Kaplan. *The Assessment of Aphasia and Related Disorders*. Lea & Febiger, 1972.

Angela Guarino, Francesca Favieri, Ilaria Boncompagni, Francesca Agostini, Micaela Cantone, and Maria Casagrande. Executive functions in alzheimer disease: A systematic review. *Front. Aging Neurosci.*, 10:437, 2018.

Mark Andrew Hall. Correlation-based feature selection for machine learning. 1999.

Shamsi T. Iqbal, Piotr D. Adamczyk, Xianjun Sam Zheng, and Brian P. Bailey. Towards an index of opportunity: Understanding changes in mental workload during task execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, page 311–320, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1581139985. doi: 10.1145/1054972.1055016. URL https://doi.org/10.1145/1054972.1055016.

Clifford R Jack, Jr, Marilyn S Albert, David S Knopman, Guy M McKhann, Reisa A Sperling, Maria C Carrillo, Bill Thies, and Creighton H Phelps. Introduction to the recommendations from the national institute on Aging-Alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimers. Dement.*, 7(3):257–262, May 2011.

Sweta Karlekar, Tong Niu, and Mohit Bansal. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models, 2018.

Heechel Kim, Hong-Woo Chun, Seonho Kim, Byoung-Youl Coh, Oh-Jin Kwon, and Yeong-Ho Moon. Longitudinal Study-Based dementia prediction for public health. *Int. J. Environ. Res. Public Health*, 14(9), August 2017.

Gina Kolata. For Scientists Racing to Cure Alzheimer's, the Math Is Getting Ugly. *NY Times*, July 2018.

Weirui Kong, Hyeju Jang, Giuseppe Carenini, and Thalia Field. A neural model for predicting dementia from language. In *Machine Learning for Healthcare Conference*, pages 270–286, 2019.

Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. Predicting confusion in information visualization from eye tracking and interaction data. In *IJCAI*, pages 2529–2535, 2016.

Yuying Liang, Yoni Pertzov, Jennifer M Nicholas, Susie M D Henley, Sebastian Crutch, Felix Woodward, Kelvin Leung, Nick C Fox, and Masud Husain. Visual short-term memory binding deficit in familial alzheimer's disease. *Cortex*, 78:150–164, May 2016.

Michael R MacAskill and Tim J Anderson. Eye movements in neurodegenerative diseases:. *Curr. Opin. Neurol.*, 29(1):61–68, February 2016.

Pascual Martínez-Gómez and Akiko Aizawa. Recognition of understanding level and language skill using measurements of reading behavior. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 95–104, 2014.

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

M-Marsel Mesulam, Emily J Rogalski, Christina Wieneke, Robert S Hurley, Changiz Geula, Eileen H Bigio, Cynthia K Thompson, and Sandra Weintraub. Primary progressive aphasia and the evolving neurology of the language network. *Nat. Rev. Neurol.*, 10(10): 554–569, October 2014.

A J Mitchell, H Beaumont, D Ferguson, M Yadegarfar, and B Stubbs. Risk of dementia and mild cognitive impairment in older people with subjective memory complaints: meta-analysis. *Acta Psychiatr. Scand.*, 130(6):439–451, December 2014.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill Education, mar 1997. ISBN 0070428077. URL https://www.xarg.org/ref/a/0070428077/.

Robert J Molitor, Philip C Ko, and Brandon A Ally. Eye Movements in Alzheimer's Disease. *J. Alzheimers. Dis.*, 44(1):1–12, January 2015.

Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.*, 53(4):695–699, April 2005.

Ivanna M Pavisic, Nicholas C Firth, Samuel Parsons, David Martinez Rego, Timothy J Shakespeare, Keir X X Yong, Catherine F Slattery, Ross W Paterson, Alexander J M Foulkes, Kirsty Macpherson, Amelia M Carton, Daniel C Alexander, John Shawe-Taylor, Nick C Fox, Jonathan M Schott, Sebastian J Crutch, and Silvia Primativo. Eyetracking

Metrics in Young Onset Alzheimer's Disease: A Window into Cognitive Visual Functions. *Front. Neurol.*, 8, August 2017.

María Luisa Barragán Pulido, Jesús Bernardino Alonso Hernández, Miguel Ángel Ferrer Ballester, Carlos Manuel Travieso González, Jiří Mekyska, and Zdeněk Smékal. Alzheimer's disease and automatic speech analysis: A review. *Expert Syst. Appl.*, 150: 113213, July 2020.

Eric M Reiman, Jessica B Langbaum, Pierre N Tariot, Francisco Lopera, Randall J Bateman, John C Morris, Reisa A Sperling, Paul S Aisen, Allen D Roses, Kathleen A Welsh-Bohmer, Maria C Carrillo, and Stacie Weninger. CAP–advancing the evaluation of preclinical alzheimer disease treatments. *Nat. Rev. Neurol.*, 12(1):56–61, January 2016.

K S Shaji, P T Sivakumar, G Prasad Rao, and Neelanjana Paul. Clinical practice guidelines for management of dementia. *Indian J. Psychiatry*, 60(Suppl 3):S312–S328, February 2018.

Reisa Sperling, Elizabeth Mormino, and Keith Johnson. The evolution of preclinical Alzheimer's disease: Implications for prevention trials. *Neuron*, 84(3):608–622, November 2014.

Reisa A Sperling, Clifford R Jack, and Paul S Aisen. Testing the Right Target and the Right Drug at the Right Stage. *Sci. Transl. Med.*, 3(111):111cm33, November 2011.

Chin Hong Tan, Bradley T Hyman, Jacinth J X Tan, Christopher P Hess, William P Dillon, Gerard D Schellenberg, Lilah M Besser, Walter A Kukull, Karolina Kauppi, Linda K McEvoy, Ole A Andreassen, Anders M Dale, Chun Chieh Fan, and Rahul S Desikan. Polygenic hazard scores in preclinical alzheimer disease. *Ann. Neurol.*, 82(3):484–488, September 2017.

A B Tobii. Specifications for the tobii eye tracker 4C, 2017.

Dereck Toker, Cristina Conati, Sébastien Lallé, and Md Abed Rahman. Further Results on Predicting Cognitive Abilities for Adaptive Visualizations. *IJCAI*, pages 1568–1574, 2017.

Dereck Toker, Cristina Conati, and Giuseppe Carenini. Gaze analysis of user characteristics in magazine style narrative visualizations. *User Modeling and User-Adapted Interaction*, 29(5):977–1011, 2019.

Laszlo Toth, Ildiko Hoffmann, Gabor Gosztolya, Veronika Vincze, Greta Szatloczki, Zoltan Banreti, Magdolna Pakaski, and Janos Kalman. A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. *Curr. Alzheimer Res.*, 15(2):130–138, 2018.

James C Vickers, Stan Mitew, Adele Woodhouse, Carmen M Fernandez-Martos, Mathew T Kirkcaldie, Alison J Canty, Graeme H McCormack, and Anna E King. Defining the earliest pathological changes of alzheimer's disease. *Curr. Alzheimer Res.*, 13(3):281–287, 2016.

Nahid Zokaei and Masud Husain. Working memory in alzheimer's disease and parkinson's disease. *Curr. Top. Behav. Neurosci.*, 41:325–344, 2019.

Stuart M Zola, C M Manzanares, P Clopton, J J Lah, and A I Levey. A behavioral task predicts conversion to mild cognitive impairment and alzheimer's disease. *Am. J. Alzheimers. Dis. Other Demen.*, 28(2):179–184, March 2013.

Appendix A. Cohort Characteristics

		Full dataset		Clean dataset	
		Patient	Control	Patient	Control
Total number of participants	N	68	73	58	62
*Participant sex	Available data (n)	68	73	58	62
	Male	34	22	27	19
	Female	34	51	31	43
*Age at enrollment	Average	71.6	64.9	71.8	64.6
	Range	52-96	50-83	52-96	50-83
	Standard deviation	9.26	9.93	9.28	9.98
*Total years of education	Available data (n)	67	72	58	62
	Average	14.8	15	14.5	14.9
	Range	9-23	12-20	9-23	12-20
	Standard deviation	3.36	2.25	3.15	2.23
*Medical history	Available data (n)	66	73	56	62
	Previous stroke	7	2	6	2
	Parkinsons disease	0	0	0	0
	Other neurological diagnosis	0	2	0	2
	History of major depression	6	8	5	7
	Other psychiatric condition	5	8	4	7
	Previous concussion	18	17	14	15
	Hypertension	27	22	22	18
	Hyperlipidemia	15	17	13	14
	Epilepsy	1	1	1	0
	HIV positive	1	1	1	1
	REM sleep disorder	2	0	2	0
	Sleep impaired	28	22	18	23
	Family history of dementia	35	33	32	29
*Tobacco & alcohol intake	Available data (n)	65	72	55	61
	Currently smoking	3	7	4	5
	Formerly smoked	24	27	23	21
	$\geq$ 4 alcoholic drinks a week	13	16	12	14
	2-4 alcoholic drinks a week	17	17	12	16
	$\leq$ 2-4 alcoholic drinks week	15	24	13	19
	Never drinks alcohol	20	15	18	12
Expert clinician diagnosis	Available data (n)	68	73	58	62
	AD	34	0	31	0
	MCI	22	0	17	0
	SMC	7	0	6	0
	Mixed Dementia†	5	0	4	0
Cookie Theft task	Available data (n)	68	73	58	62
completion time (secs.)	Average	67.4	86.1	69.8	82.0
	Range	19-257	12-357	19-258	20-357
	Standard deviation	40.2	67.0	41.9	62.5
Functional dependency	Available data (n)	66	73	56	62
	Independent	48	72	39	61
	Dependent	18	1	17	1
Sleep routine (h/night)	Available data (n)	68	73	58	62
	Average	7.6	7	7.7	7.0

	Range	3-12	3-11	5-12	3-11
	Standard deviation	1.54	1.33	1.55	1.34
MoCA test results (0-30 scale)	Available data (n)	59	71	50	60
	Average	20.25	27.15	20.2	27.1
	Range	5-29	19-30	5-29	19-30
	Standard deviation	5.44	2.73	5.51	2.72
ApoE status	Available data (n)	29	0	28	0
	e2e3	1	-	1	-
	e3e3	9	-	9	-
	e3e4	16	-	16	-
	e4e3	1	-	1	-
	e4e4	2	-	1	-
Subjective memory complaint (0-13 scale)	Available data (n)	45	41	38	33
	Average	6.42	2.32	6.37	2.15
	Range	0-13	0-7	0-13	0-7
	Standard deviation	3.19	1.60	3.19	1.63
First language learned	Available data (n)	66	73	56	62
	English	58	60	50	50
	Other	8	13	6	12
Preferred language to speak	Available data (n)	67	72	57	61
	English	63	69	54	58
	Other	4	3	3	3
Self-identified origin(s)	Available data (n)	67	73	57	62
	North American Indigenous	6	0	6	0
	Other North American	29	26	25	26
	European	41	46	41	39
	Caribbean	1	0	1	0
	Latin America	1	0	1	0
	African	1	1	1	1
	Middle Eastern	1	1	1	1
	South Asian	1	1	1	0
	East Asian	3	6	2	5
	Other	1	3	1	2

†characteristics of AD and vascular dementia

* features comprising the demographics baseline

Appendix B. Experience with the Technology Questionnaire

Experience with the Technology	Patient (N=56)	Control (N=56)
Was comfortable during the assessment	93%	95%
Had privacy concerns using this technology	5%	2%
Was relaxed during the assessment	91%	98%
Was engaged and interested during the assessment	93%	98%
Willing to repeat the assessment on a yearly basis	94%	96%
Willing to repeat the assessment on a monthly basis	42%	50%
Willing to repeat the assessment on a weekly basis	16%	16%
Willing to repeat the assessment on a daily basis	2%	9%