Increasing access to cognitive screening in the elderly: applying natural language processing methods to speech collected over the telephone.

Catherine Diaz-Asper[a]*, Chelsea Chandler[b], Raymond Scott Turner[c], Brigid Reynolds[c], & Brita Elvevåg[d]

a Department of Psychology, Marymount University, Arlington VA, USA;

b Department of Computer Science, University of Colorado, Boulder, USA;

c Department of Neurology, Georgetown University, Washington DC, USA;

d Department of Clinical Medicine, University of Tromsø - the Arctic University of Norway, Norway

* Corresponding author:

Department of Psychology

Marymount University

2807 N. Glebe Road

Arlington, VA 22207-4299

cdiazasp@marymount.edu

ABSTRACT

Barriers to healthcare access are widespread in elderly populations, with a major consequence that older people are not benefiting from the latest technologies to diagnose disease. Recent advances in the automated analysis of speech show promising results in the identification of cognitive decline associated with Alzheimer's disease (AD), as well as its purported pre-clinical stage. We utilized automated methods to analyze speech recorded over the telephone in 91 community-dwelling older adults diagnosed with mild Alzheimer's disease (AD), amnestic mild cognitive impairment (aMCI), or cognitively healthy. We asked whether natural language processing (NLP) and machine learning could more accurately identify groups than traditional screening tools and be sensitive to subtle differences in speech between the groups. Despite variable recording quality, NLP methods differentiated the three groups with greater accuracy than two traditional dementia screeners and a clinician who read transcripts of their speech. Imperfect speech data collected via a telephone is of sufficient quality to be examined with the latest speech technologies. Critically, these data reveal significant differences in speech that closely match the clinical diagnoses of AD, aMCI and healthy control.

KEYWORDS

AD, amnestic MCI, cognitive screening, NLP, machine learning

1. INTRODUCTION

Unequal access to health services is a growing problem for the elderly, and is attributable to vast differences in financial resources, geographical location and the obvious physical challenge of attending in-person to clinics when old and frail [1]. Such inequity in terms of access to basic services thus compounds the effects of aging and results in the elderly population not being uniformly able to take advantage of proactive health monitoring services and some advances in diagnostic methods. Therefore, this study assessed cognitive function using brief conversational tasks administered via the telephone, thereby obviating the need for in-person attendance. Detecting cognitive decline as early as possible is important to enable planning for the future, increase quality of life, reduce care costs and potentially gain added benefit from therapeutic drug trials [2]. However, current screening methods typically fail to detect it until such time when decline in memory and other cognitive functions are clearly evident.

Early signs of cognitive decline may be evident in speech [3,4] such that it is possible to differentiate cognitively healthy from individuals with Alzheimer's disease (AD) via speech alone [5,6] in highly controlled settings, but it is unknown whether this generalizes to naturalistic settings. The value of remote screening could be enormous, both in terms of earlier detection and for increased access. This study sought to establish if it is possible to detect these early signs in speech in a group who are at high risk for later conversion to AD [7,8] and often remain undiagnosed due to poor sensitivity of traditional screening tests (i.e., amnestic mild cognitive impairment

(aMCI)), by leveraging natural language processing (NLP) methods on speech collected in naturalistic settings. Specifically, we asked three questions: (1) What are the clinically relevant language features that best differentiate mild AD, aMCI and healthy controls? We hypothesized that speech coherence or intelligibility would differentiate control participants from AD participants, with the aMCI group intermediate between the two [9-11]. (2) How well do NLP features and machine learning methods classify the three groups? We hypothesized that our models would be able to separate the three groups from one another, with the healthy controls most obviously different from the individuals with AD, and the aMCI group intermediate between the other two. Previous studies in controlled settings [5,6] have reported models capable of distinguishing healthy, aMCI and AD groups with good measures of separability (i.e., area under the receiver operating characteristic curve (AUC) in the 0.70 range. The higher the AUC, the better the model is at predicting who is in which group). We expected to find similar measures of separability using our real-world data, collected under significantly less controlled conditions. (3) Can automated methods provide more accurate diagnostic predictions than traditional dementia screening tools or expert humans? We hypothesized more accurate group categorizations than traditional screening tools and experts (with no contextual knowledge beyond a speech sample), due to highly sensitive machine learning techniques.

2. MATERIALS & METHODS

Here we report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to

data analysis, all manipulations, and all measures in the study. No part of the study procedures or analyses was pre-registered prior to the research being conducted.

2.1 Participants

Participants (N = 91)[1] were community-dwelling English speakers, recruited via the Memory Disorders Program (MDP) at Georgetown University (Table 1). One third carried a diagnosis of mild AD, one third amnestic MCI, and one third were cognitively healthy. Clinical diagnoses of AD [12] and mild cognitive impairment [13] followed established criteria. Since mild cognitive impairment is a heterogeneous clinical syndrome, individuals with aMCI (single or multiple domain [14]) were included to reduce clinical variability as this subgroup is at greatest risk for conversion to AD [7,8]. Control participants had no significant medical history or subjective cognitive complaint. All participants had Mini Mental State Exam (MMSE [15]) scores within the last 6 months, had adequate hearing, and no self-reported history of neurological disease (e.g., Parkinson's disease or epilepsy), drug or alcohol abuse, psychiatric hospitalization, current cancer treatment, or stroke or heart attack within the last year. Individuals with minor physical ailments (e.g., diabetes with no serious complications, essential hypertension) were included. Participant recruitment, written informed consent (with authorized representatives also providing consent for participants in the mild AD group), medical history and administration of the MMSE were conducted in the Georgetown University MDP prior to the telephone interview. Only contact details for each participant were shared with the telephone interviewer, who remained blind to

[1] An a-priori analysis indicated that, to achieve power of .80 and a moderate effect size (f^2=.40), a total sample size of at least 66 would be required to detect a significant model (f(2,63)=3.14,p<.05).

participant diagnostic group. The study was approved by the institutional review boards of Marymount University and Georgetown University (MU IRB#260). All inclusion/exclusion criteria were established prior to participant recruitment.

INSERT TABLE 1 HERE

2.2 Materials & Procedures

The telephone interview (approx. 20 mins) (i) collected speech samples and (ii) administered a screening test for cognitive decline (in counterbalanced order).

Telephone Screener: A modified version of a telephone based screening instrument for cognitive decline - the Telephone Interview for Cognitive Status [16] - was employed (TICS-M). The TICS-M is modeled after the MMSE in providing a brief, global measure of cognitive functioning, and has good sensitivity and specificity to detect dementia [17], but its utility to screen for milder cognitive syndromes is unknown [18,19]. Legal copyright restrictions prevent public archiving of the TICS-M and MMSE which can be obtained from the copyright holders in the cited references.

Speech Samples: Participants generated as many 'animal' words as possible in one minute (semantic word fluency), and described a favorite memory from childhood (free speech) (Table 1).

Participants were telephoned at home via the Cisco Jabber interface on a laptop computer, and the semantic word fluency and free speech portions of the interview were recorded onto the device and later uploaded to a secure cloud-based application. Spouses/companions were asked to remove visual memory aids (e.g., calendars) and turn off audible distractors prior to the interview. The speech samples were digitally

recorded and transcribed by the first author or a trained research assistant (intraclass correlation coefficient = 0.988) to check for accuracy and screen for personally identifying information. (The conditions of our ethics approval do not permit public archiving of anonymized study data. Readers seeking access to the data should contact the corresponding author. Access will be granted to named individuals in accordance with ethical procedures governing the reuse of sensitive data. Specifically, requestors must meet the following conditions to obtain the data: completion of a formal data sharing agreement; approval by the Marymount and Georgetown University IRBs).

2.3 Data Analysis

A range of natural language features were extracted from participant responses in the free speech task and the semantic fluency task. In general, features were extracted automatically using custom written Python code and various packages for data management, statistical calculations, NLP analyses, and word vector creation. For each classification setting, the most predictive and clinically relevant features were chosen to train and test machine learning models. The best performing models are reported in the Results.

2.3.1 NLP Feature Selection

For the *free speech task*, three classes of NLP features (a set of 73 total) were extracted, namely (i) word-level (lexeme), (ii) sentence-level (syntactic), and (iii) meaning (semantics) of expressions [20]. The first class of language features included simple counts of word tokens and word types (i.e., unique words), and slightly more

sophisticated metrics (type to token ratio (TTR; a measure of lexical richness), content density (a measure of actual information spoken, as opposed to filler words), Brunét's Index (a measure of lexical richness less affected by text length), Moving Average Type Token Ratio (a version of TTR that is calculated on a sliding window of the text and is less affected by text length) Honoré's Statistic (emphasizes words that are only spoken once), and counts and frequencies of specific parts of speech that were computed with NLTK's standard TreeBank tagger (https://www.nltk.org/). Some of these features are more impacted by text length as longer utterances will receive a higher score. Since poverty of speech is a common symptom in conditions such as AD, features that take this into account tend to be more highly discriminable of the AD group than those that do not have such an effect.

The second class of language features were syntactic features, or those that seek to measure the complexity and arrangement of sentences. These included measures extracted from dependency parses or speech graphs. Examples of such metrics are distances of dependencies in parses or the number of nodes, edges, or loops in speech graphs.

The third class, semantic features, were computed in a few different ways. Generally, semantic analyses are performed using high-dimensional vector space word embeddings of text. These embeddings operate under the premise that the meaning of a word is derived from the context in which it tends to appear. Words that tend to appear in similar contexts are semantically related and thus should be close to each other in a derived vector space. Examples of word embedding techniques are Latent Semantic Analysis (LSA; [21]), word2vec [22], Embeddings from Language Models

(ELMo; [23]), and Bidirectional Encoder Representations from Transformers (BERT; [24]). LSA performs a singular value decomposition on a sparse type-to-document matrix to obtain lower dimensional vectors of each of the types. Word2vec is a neural network-based word embedding model trained on a large corpus of text with the goal of predicting either a word given its context or the context surrounding a word given the word. ELMo and BERT are deep neural language models that are built on long short term memory neural language models and transformers, respectively. Metrics are computed on the cosine distances between consecutive embeddings or windows of embeddings, or by calculating the slope of coherence through the text. For end-to-end models like BERT, the entire network can be harnessed and subsequently tuned with a new layer to produce predictions.

For the *semantic fluency task*, a task-specific feature set (comprising 26 features) was extracted from participant responses. Traditionally, the semantic fluency task is administered and scored by trained humans who count the number of unique items (in this case, animals) spoken. More detailed analyses of responses to this task have been proposed by researchers that can provide additional insights into human cognitive performance [25-28]. Classically, Troyer et al. [29] proposed two metrics that measure important components of the animal fluency task - clustering (i.e., producing words within the same subcategory of animals, like *safari animals* or *house pets*) and switching (i.e., changing between clusters). This approach can be implemented with hand-coded categories of animals or by using semantic distances. Using semantic distances entails computing the cosine distance of the word embedding of each exemplar to the next and setting a threshold of belonging to a category or not.

A number of features were extracted from the semantic fluency task, namely the number of unique animals spoken, the number of categories produced (employing both the hand-coded Troyer categories as well as a BERT [24] word embedding-based thresholding method where cosine distances between consecutive BERT representations of animal words are computed and those distances that fall below a predetermined threshold are considered a jump to a new category of animals), the average number of animals per category, the average cosine similarity between successive animals and successive categories, and the average vector length of each exemplar's word embedding.

Each time the average was computed, so too was the standard deviation, minimum, and maximum. The length of the animal vectors has been shown previously to be an indicator of the "usualness" of the animals spoken [30]. Researchers in NLP have shown that words that occur in many different contexts, and thus have less meaning (such as stop words or other commonly used words), move around in vector space during computation and are shortened with each move due to an averaging computation. Thus, the longer the vector representation is, the more unusual the word tends to be [31].

The discriminability of each feature was determined by multivariate statistical analyses (specifically f-statistics) and feature importance in machine learning models. Specifically, the NLP features with an f-statistic greater than 5.0 (range of f-statistics for features in all prediction scenarios: free speech, 0.00 - 11.72; animal fluency, 0.00 - 35.01) were initially chosen for experimentation and the machine learning models further narrowed down this choice by eliminating those features that were not critical for

increasing model performance (e.g., due to multicollinearity with other features). Features were computed for the entire dataset, but each prediction setting followed its own distinct feature selection process on its corresponding labels and data.

2.3.2 Classification

We sought to answer how well contemporary NLP methods can differentiate the three groups and whether machine learning methods can inform further about the relative importance of language variables in different stages of decline. When performing these experiments, some models overfitted the samples' idiosyncratic characteristics such that some features were statistically important in differentiating groups, but lacked clinical relevance (e.g., amount of numbers used in free speech) and thus were omitted to improve potential generalizability.

For each classification setting, we first performed a feature selection process that narrowed down our feature set to those that had the highest discriminability, yet were also clinically relevant (detailed in the first section of the Results). Then we used a grid search methodology to optimize the hyperparameters, and investigated 7 different machine learning model architectures (specifically a Decision Tree Classifier, Extra Trees Classifier, Gradient Boosting Classifier, K Neighbors Classifier, Logistic Regression Classifier, Random Forest Classifier, and Support Vector Classifier), including those with 0-4 tunable hyperparameters with 1-13 options each. The grid search was performed with the goal of not just building the best model, but rather understanding the relevance of features and how they may be used for the detection of dementia. If certain features were consistently implicated in each model, it would be

clear that they were not simply idiosyncratic to a particular algorithm. Furthermore, it was important to explore which model and hyperparameter combinations tended to work best with the distribution of the chosen features. Decision Tree Classifiers, Extra Trees Classifiers and Gradient Boosting Classifiers worked particularly well, and placed consistently in the top 10% of model architectures for each scenario, as they assume no prior distribution of the data, do not depend on probability distribution assumptions, and allow the data to be partitioned on different combinations of the chosen features. They also tend to have excellent accuracy with high-dimensional datasets.

In the Results sections, we report statistics of the accuracy of not only the top performing model, but also the top 10% of the models tested so as to offer transparency around the level of accuracy consistency in the overall results of the grid search. We used leave-one-out cross-validation in each setting as this type of cross-validation allowed us to simulate how the model would predict a new participant after being fully trained on our initial dataset.

Codes for feature extraction and model training can be accessed at:

https://github.com/ckchandler/Increasing-Access-to-Cognitive-Screening

3. RESULTS

3.1 What are the clinically relevant language features that best differentiate between the three groups?

Aberrations in meaning and language have been identified as critical indicators of cognitive decline in both aMCI and AD [9,19], thus we focused on text-based analyses. All NLP features with significant f-values in group comparisons for both the free speech task and the semantic fluency task are listed in Table 2.

For the *free speech task*, certain word-level features consistent with poverty of speech (*raw count* of nouns, determiners, present participle verbs, and modals) had statistically significant f-values when comparing the AD group to the cognitively healthy group and to the aMCI group. (A modal is a type of verb that is used to indicate modality such as likelihood, requests, suggestions, and so on – for example, *can*, *could*, *may*, and *might*; the frequency is computed by dividing the number of modals spoken by the total number of words spoken).

Other word-level features (*frequency* of modals, past participle verbs, non third person singular verbs, and all verb types) had statistically significant f-values only when comparing cognitively healthy participants to those with an aMCI diagnosis. For syntactic features, the mean distance of all dependencies between words in a sentence in a participant response served as a discriminable feature for the AD group when comparing to both the cognitively healthy group and the aMCI group, but did not significantly differentiate the cognitively healthy group from the aMCI group. The semantic feature that proved to be most discriminable in our dataset was the mean coherence of a 4 word sliding window of the 300-dimensional word2vec word embeddings based on 3 million words from the Google News corpus. The window size (4 words) is a hyper-parameter that is generally tuned to be whatever size produces the most accurate representation of pieces of text; at a high level, each window should

represent a distinct phrase so as to smooth out the noise that would be produced if comparing consecutive words. We found that this feature discriminated the AD group from both the cognitively healthy group and the aMCI group, but failed to do the same with discriminating the aMCI group from the participants labeled as cognitively healthy. For the *semantic fluency task*, the number of unique animals spoken was the most discriminable feature overall; it was the highest for separating AD from cognitively healthy, fairly high when comparing aMCI to AD, and less high - yet still significant - for separating cognitively healthy from aMCI. The same can be said for the number of categories, based on the hand-coded Troyer et al. [29] categories, but with slightly less discriminatory power. Finally, with even less discriminatory power, yet still a significant amount, the maximum number of animals spoken per category (i.e., the maximum number of animals spoken consecutively from one category), discriminated all groups fairly well.

INSERT TABLE 2 HERE

3.1.1 Task Specificity

Another finding of the feature extraction portion of our work is that there is no single language feature that is consistent between differing tasks that may be discriminable for varying levels of cognitive decline. As an example, we discuss coherence in language and how it varies between tasks. Previous work by our team has found that coherence in recalling a short story is generally lower in a group of individuals with varying levels of mental illness (see [32] and [33] for an overview of this work) and that higher coherence in story recalls generally received higher expert ratings of recall as well. In the current

study, we found the opposite: that *lower* coherence actually belonged to the cognitively healthy group, then the aMCI group, and finally the AD group had the highest cohesion. The methodology was the same in both approaches - the average cosine distance between consecutive windows of size 4 was computed for each response. There certainly will be differences when coherence is operationalized in different ways (see [34] for an overview of different approaches to computing coherence), but this is not a factor here. The only difference between the two experimental settings was the task. One is a constrained task where the participants try to remember specific details of a short story recently told to them, and the other is free speech where the response is given in a narrative manner, relying on long-term autobiographical memory, and likely retold with greater emotion and enthusiasm. The higher coherence of the AD group in this dataset could be attributed to more repeated words and less detail overall. To illustrate, we include portions of text from a participant from the AD group with the highest computed coherence to show how a wordy response that consists of repeated statements would generate a high coherence:

> *“...And uh, it wasn’t a project actually it was a it was a um. It, it was a, it was a house. Um. It was a not a house, it was a it was a um, a development… She, she never learned to read and write. Um, but she um uh, I don’t, I don’t think she ever learned to read and write, but she um, uh she may have. I, I think she may have learned to, to read and write..."*

Since this phenomenon is between two different studies and thus two different participant pools, we also explored whether the coherence between individual exemplars in participants' animal fluency responses correlated with the coherence of participants' childhood memory response. As noted earlier, these are two dissimilar tasks, tapping quite different cognitive processes, and it is perhaps unsurprising that we found a low correlation between the two features (Pearson *r* correlation of 0.26). Thus, we conclude that there is little commonality across tasks and advocate for task-specific measures and methods of computing such measures (e.g., either using larger window sizes to account for long, drawn out phrases, or removing verbatim repeated clauses in the case that repeated words and more verbosity in general are expected). We further advocate here that researchers working with computational methods *must* be explicit in reporting the manner in which their metrics were computed, especially in the cases where there do not yet exist standardized methodologies.

3.2 How well do language features and machine learning methods classify the three groups?

3.2.1 Classifying Cognitively Healthy, aMCI, and AD

In the setting of classifying the three groups together as an assay of level of cognitive decline, the top three features chosen for machine learning modeling were the average coherence in free speech, and the number of unique animals and categories spoken in semantic fluency.

We used these three features to train a model for classifying cognitively healthy, aMCI, and AD participants. The best model in this experimental setting was a Decision Tree Classifier with a maximum tree depth of 3. This model was 62% accurate overall when performing leave-one-out cross-validation (Table 3; Appendix A). Figure 1 shows the ROC curve of each of the three groups.

INSERT FIGURE 1 & TABLE 3 HERE

The model was most accurate at predicting cognitively healthy, then aMCI, and was least accurate in the AD setting.

As a method to visualize the diagnostic groups based solely on these three features, we applied Principal Component Analysis (PCA) to the data. Figure 2 (top left panel) shows density plots of the first dimension of this reduction, separated by diagnosis. The distributions are ordered as expected, with the cognitively healthy and AD groups at the two extremes. The left edge of the aMCI peak aligned with the peak of cognitively healthy group, while the right edge of the aMCI peak aligned with the peak of the AD group. If an aMCI participant was incorrectly predicted as cognitively healthy, that is because they were within the healthy range for this sample. Similarly, if an aMCI participant was incorrectly predicted as a member of the AD group, that is because they performed more within the AD range. For these individuals, this prediction “error” could signpost a future conversion to AD.

Of the models tested, the top 10% of the models were on average 53% accurate (SD 0.03, minimum 51%, maximum 62%).

3.2.2 Classifying Cognitively Healthy against "Cognitive Decline"

Next, we tested the setting of classifying the cognitively healthy group against cognitive decline in general (i.e., aMCI and AD participants were treated as belonging to the same group). Since coherence was not a significant indicator for differentiating cognitively healthy from aMCI, in this setting, we replaced coherence with the frequency of modals in the language. This feature, plus the number of unique animals spoken and the number of categories spoken were used to train a machine learning classifier to classify cognitively healthy against cognitive decline. The best model found in this experimental setting was an Extra Trees Classifier (a classifier comprising 32 Decision Trees) with a maximum tree depth of 10 and an entropy criterion for separating the data into subsets with more homogeneity within individual groups (Table 3; Appendix A). This model was 87% accurate overall when performing leave-one-out cross-validation and had an AUC of 0.86.

Of the models tested, the top 10% of the models were on average 84% accurate (SD 0.02, minimum 81%, maximum 87%).

3.2.3 Classifying Cognitively Healthy and aMCI

The next machine learning model implemented was that of distinguishing cognitively healthy from aMCI. This setting is where the dichotomy between the most accurate model and the most clinically relevant model was apparent. When allowing the machine learning model to choose the best features to differentiate the two groups, it overwhelmingly favored features such as cardinal (number) counts and frequencies, the

frequency of non-third person singular verbs and wh-adverbs (e.g., *when, where, why*), and the number of unique animals spoken in semantic fluency. This model achieved an accuracy of 87% and an AUC of 0.88, which is high compared to other studies (e.g., [6]). These features are not backed by literature or clinical relevance, so we report another model that, while less accurate on this dataset, has the potential to be more generalizable and have greater translational value.

The clinically relevant model for this comparison was one based only on the number of unique animals and categories generated in the fluency task. Interestingly, the clinically relevant features extracted for the cognitively healthy and aMCI from the free speech task did not add additional information that was distinct from the differences derived from the fluency task. The best model in this setting was a Decision Tree Classifier with a maximum tree depth of 4. It achieved an accuracy of 80% and an AUC of 0.78 (Table 3; Appendix A). The model correctly predicted 86% of the cognitively healthy group and 75% of the aMCI group. Figure 2 (top right panel) shows the PCA dimensionality reduction of these two groups performed on their top features as determined by the f-value.

Of the models tested, the top 10% of the models in the clinically relevant feature setting were on average 76% accurate (SD 0.02, minimum 71%, maximum 80%).

3.2.4 Classifying Cognitively Healthy and AD

The best model for differentiating cognitively healthy from AD was based on the number of turns spoken by the participant in the free speech task (broken up by prompts to

continue speaking by the interviewer; some participants had one turn, but others needed to be asked many follow-up questions to continue talking), and the number of repeated animals, categories, and the average and maximum length of the animal word vectors spoken during the semantic fluency task. The best model was Extra Trees Classifier with 32 estimators, a maximum tree depth of 5, and an entropy criterion. It achieved an accuracy of 88% and an AUC of 0.90 (Table 3; Appendix A). Out of the 29 cognitively healthy participants, 2 were predicted as AD and of the 30 AD participants, 5 were predicted as cognitively healthy. Figure 2 (bottom left panel) shows the PCA dimensionality reduction of the data used in the machine learning model.

Of the models tested, the top 10% of the models were on average 87% accurate (SD 0.01, minimum 87%, maximum 88%).

3.2.5 Classifying aMCI and AD

Finally, we discuss the setting of differentiating aMCI from AD. This is of critical interest for clinical translational value as those with aMCI are at an increased risk to convert to AD. Thus, incorrect predictions for the aMCI group may indicate people who are more likely to convert to AD. A Decision Tree Classifier with a maximum depth of 3 based on the mean coherence in free speech and the number of unique animals, categories, and maximum coherence between successive animals in semantic fluency resulted in an impressive 79% accuracy and 0.74 AUC (Table 3; Appendix A). Figure 2 (bottom right panel) shows the PCA dimensionality reduction of these two groups alone performed on their top features as determined by the f-value.

INSERT FIGURE 2 & TABLE 4 HERE

Of the models tested, the top 10% of the models were on average 75% accurate (SD 0.03, minimum 69%, maximum 79%).

3.3 Can natural language processing models provide more accurate diagnostic predictions than traditional dementia screening methods?

3.3.1 Comparison to human judgement

Blind to diagnosis, co-author R.S.T., a neurologist specializing in the diagnosis of dementia, labeled the transcript of each participant's free speech response as belonging to one of the three groups. The resulting labels assigned to each participant were 49.45% accurate (Appendix B). We present two comparisons of the human classification to our machine learning models. The first used our best machine learning model (based on both the free speech task and the fluency task; labeled ML in Figure 3), whereas the second used a separate model based only on the free speech task (labeled ML (fs) in Figure 3) to more accurately compare to the resources available for R.S.T's labeling.

The human classifications were less accurate than our *best* model in predicting who was cognitively healthy (58.6% accurate versus 75.8% in the machine learning model) and aMCI (34.4% accurate versus 65.6% in the machine learning model). However, the human was more accurate in predicting AD than the machine learning model (56.7% accurate versus 43.3% in the machine learning model).

When comparing the human classifications with the machine learning model *based only on the free speech data*, R.S.T was more accurate in identifying cognitively healthy participants than our free speech based model (24.1%), but less accurate for aMCI, (46.87%) and AD (63.3%). This suggests that the free speech portion of our testing battery held the most signal in differentiating aMCI from AD, but very little signal in classifying cognitively healthy. The fluency portion carried much of the diagnostic accuracy in identifying cognitively healthy, but less so in the classes with higher degrees of cognitive impairment.

INSERT FIGURE 3 HERE

3.3.2 Comparison to traditional screening tools

Several different dementia screening tools are used in clinical practice, but here we report on one common and widely-used paper-and-pencil screener (the Mini-Mental State Examination, MMSE; [15]) and a similar screener designed to be administered over the telephone (the modified Telephone Interview for Cognitive Status, TICS-M; [16]). We compared our automated approach to the MMSE and TICS-M in their ability to classify diagnostic groups, and found that our approach generally outperformed the traditional screening tests. Furthermore, the MMSE categorization had a stronger correlation with participant education level (correlation between education and group label: $r = 0.271$, $p = 0.010$). The TICS-M and the machine learning model categorizations were not significantly correlated with education ($r = 0.093$, $p = 0.387$ and $r = 0.095$, $p = 0.375$, respectively).

The Modified Telephone Interview for Cognitive Status

We used the Knopman et al. [18] thresholds to differentiate our three groups. The point cutoffs for the AD group is between 0 and 27 (inclusive), the aMCI group is between 27 and 31 (inclusive), and then the cognitively healthy group is any score above 31. The confusion matrix for this cutoff on our participants' scores is contained in Appendix B. This shows a high skew towards cognitively healthy predictions.

As our sample was highly educated, we sought to control for this effect on predictions with an education scaling factor. Following the guidelines set forth by Knopman et al. [18], we adjusted for participant education level by not adding any points to the raw TICS-M score for participants with between 11 and 15 years of education and subtracting 2 points for subjects with 16 or more years of education. Surprisingly, accuracy declined for the AD group when scaling for education as three previously correctly predicted AD participants were mislabeled into the aMCI group. The cognitively healthy and aMCI groups were unaffected.

Our best machine learning model for classifying cognitively healthy, aMCI, and AD was more accurate than the TICS-M test. The TICS-M test was 45% accurate on our participants, even when scaled for education effects. The TICS-M test overwhelmingly classified the participants as cognitively healthy, whereas there was a more even spread of predictions in the machine learning approach.

Mini-Mental State Examination

We again explored various cutoffs for the MMSE and chose to report the best performing cutoff on our dataset, which was set forth by Chapman et al. [35]. (The AD group was between 0 and 24 (inclusive), the aMCI group was between 25 and 26 (inclusive), and the cognitively healthy group was any score above 27.

Appendix B contains the confusion matrix using this cutoff for the three groups. The MMSE test was 42% accurate on our participants using the Chapman et al. [35] cutoffs. Our machine learning classification model achieved higher accuracy than the MMSE on our participants.

4. DISCUSSION

Reliably recognizing cognitive decline at the earliest stage is difficult [36, 37], yet recent advances in NLP and machine learning have made considerable progress in this regard. The ultimate goal is to translate this progress into a screening and monitoring tool that can facilitate equal access for older people with memory concerns, regardless of their location, mobility status and so on. While some studies (e.g., [5]) have reported impressive accuracy in the detection of cognitive decline using these technologies, the vast majority of these have recorded speech in controlled, laboratory-based settings, which may not generalize well to the elderly living in the community who are unwilling or unable to attend an in-person clinical evaluation. Using low cost methods, the current study demonstrated that NLP and machine learning could be successfully applied to

speech of variable quality, recorded remotely from telephone conversations conducted from participants' homes.

We do note however, that the participants in our study were recruited from a database of research volunteers who were relatively homogeneous in terms of important demographics such as race and education (being predominantly White and highly educated). This is a widespread characteristic of clinical research [38], which notably limits the generalizability of findings. Furthermore, machine learning models tend to learn from and propagate societal biases between demographic groups [39] and off-the-shelf NLP models themselves are known to inflate disparities [40]. This is a critical issue that the entire field is facing [41], thus we are careful not to make generalization claims and acknowledge that further widespread work must be done to decrease this inequity.

Another limitation of this study is the small sample size. In many computational studies in the clinical domain, only rarely is there a dataset large enough to have a separate evaluation set distinct from the training and validation sets used to train the models [42]. In our case, we chose to implement leave-one-out cross-validation as these results will be the best estimate of how the model would behave when fully trained and applied to new data. Although this sort of cross validation is a common approach with small datasets (e.g., [6] [43]), we do note that it has issues with overfitting and variable test results.

Our machine learning predictions, based on a small amount of speech data recorded in suboptimal conditions, showed strong AUCs for classifying groups, and outperformed

the judgements of an expert clinician and two traditional screening tests. Our intent with these analyses was not to pit “human versus machine,” but rather to demonstrate that a machine learning approach can detect subtle diagnostic differences from just small samples of speech, and as such, could be a potential *adjunct* screener in a clinician’s battery to reach those individuals who might not otherwise be seen. As with traditional screening tools, a concerning result would signal the need for a comprehensive dementia workup.

Despite the proof of concept for this approach, much remains to be done. The next steps are to measure the predictive ability of our models to identify at an early stage who will go on to decline over time. Of particular interest in these analyses would be the cognitively healthy participants identified in our current models as belonging to the aMCI group, and the aMCI individuals identified as belonging in the AD group. If these participants were indeed to subsequently display cognitive decline and convert to aMCI and AD respectively, then the predictive ability of our models would be confirmed. Further, it remains unclear how well these models will perform with different speech and written language tasks, and this has important implications for future protocol development and to answer the question of whether a single *ideal* task can elicit the most accurate predictions. Finally, to address possible etiological heterogeneity in participant groupings, future studies would be strengthened by comparing clinical diagnostic categories against validated biomarkers of disease.

A significant advantage of our approach (when compared to traditional screening tests) is that speech can be recorded countless times without the confounding influence of practice effects or interrater variability. Hence, in future work, intra-individual variability should be measured via repeated testing at various time intervals to tease apart the effects of comorbidities, medications and the like on cognition. By demonstrating that state of the art automated methods can successfully be applied to suboptimal speech data, we address both the issue of early identification of cognitive decline, and accessibility of health care. Consequently, we are one step closer to the development of a remote, low cost, sensitive and highly accessible tool for cognitive screening on a large scale.

ACKNOWLEDGEMENTS

The authors would like to thank the study participants for their time and Kelly Ha for her support as a research assistant.

Funding: This work was supported by the National Institutes on Aging [grant number R03AG052416].

CONFLICT OF INTEREST/DISCLOSURE STATEMENT

The authors have no conflict of interest to report

REFERENCES

[1] United Nations. Health Inequalities in Old Age. 2018. https://www.un.org/development/desa/ageing/news/2018/04/health-inequalities-in-old-age/

[2] B. Dubois, A. Padovani, P. Scheltens, A. Rossi, G. Dell'Agnello. Timely diagnosis for Alzheimer's disease: a literature review on benefits and challenges. Journal of Alzheimer's Disease 2016;49:617-631 doi: 10.3233/JAD-150692.

[3] V. Berisha, S. Wang, A. LaCross, J. Liss, Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case study comparing the press conferences of presidents Ronald Reagan and George Herbert Walker Bush. Journal of Alzheimer's Disease 2015;45:959-963 doi: 10.3233/JAD-142763.

[4] R. Filiou, N. Bier, A. Slegers, B. Houzé, P. Belchior, S. Brambati, Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: a scoping review. Aphasilogy 2020;34:723-755 doi: https://doi.org/10.1080/02687038.2019.1608502.

[5] E. Eyigoz, S. Mathur, M. Santamaria, G. Cecchi, M. Naylor, Linguistic markers predict onset of Alzheimer's disease. EClinicalMedicine 2020;28 doi: https://doi.org/10.1016/j.eclinm.2020.100583.

[6] S. Orimaye, J. Wong, C. Wong, Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. PLoS ONE 2018;13:e0205636 doi: https://doi.org/10.1371/journal.pone.0205636.

[7] T.L. Michaud, D. Su, M. Siahpush, D.L. Murman, The risk of incident mild cognitive impairment and progression to dementia considering mild cognitive impairment subtypes. Dement Geriatr Cogn Disord Extra 2017;7:15-29 doi: 10.1159/000452486.

[8] K. Schmidtke, S. Hermeneit, High rate of conversion to Alzheimer's disease in a cohort of amnestic MCI patients. International Psychogeriatrics 2008; 20:96-108 doi: 10.1017/S104161020700550.

[9] S. Pakhomov, D. Chacon, M. Wicklund, J. Gundel, Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. Behavior Research Methods 2011;43:136-144 doi: https://doi.org/10.3758/s13428-010-0037-9.

[10] C. Toledo, S. Aluísio, L. dos Santos, S. Brucki, E. Trés, M. de Oliveira, *et al*. Analysis of macrolinguistic aspects of narratives from individuals with Alzheimer's disease, mild cognitive impairment, and no cognitive impairment. Alzheimer's & Dementia 2017; 10:31-40 doi: 10.1016/j.dadm.2017.08.005.

[11] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, I. Hoffmann, Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. Comput. Speech Lang. 2019; 53:181-197 doi: https://doi.org/10.1016/j.csl.2018.07.007.

[12] G. McKhann, D. Knopman, H. Chertkow, B. Hyman, C. Jack, C. Kawas, C, *et al*. 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 2011;7:263-269 doi: 10.1016/j.jalz.2011.03.005.

[13] R. Petersen, G. Smith, S. Waring, R. Ivnik, E. Tangalos, E. Kokmen, Mild cognitive impairment: clinical characterization and outcome. Arch. Neurol. 1999;56:303-308 doi: 10.1001/archneur.56.3.303.

[14] R. Petersen, J. Morris, Mild cognitive impairment as a clinical entity and treatment target. Arch. Neurol 2005;62:1160-1163 doi: 10.1001/archneur.62.7.1160.

[15] M. Folstein, S. Folstein, P. McHugh, “Mini-mental state.” A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 1975;12:189-198 doi: 10.1016/0022-3956(75)90026-6.

[16] J. Brandt, M. Spencer, M. Folstein, The Telephone Interview for Cognitive Status. Cognitive and Behavioral Neurology 1988;1:111-117.

[17] J. Gallo, J. Breitner, Alzheimer's disease in the NAS-NRC Registry of ageing twin veterans: IV. Performance characteristics of a two-stage telephone screening procedure for Alzheimer's dementia. Psychological Medicine 1995;25:1211-1219 doi: 10.1017/s0033291700033183.

[18] D. Knopman, R. Roberts, Y. Geda, V. Pankratz, T. Christianson, R. Petersen, *et al*. Validation of the Telephone Interview for Cognitive Status-modified in subjects with normal cognition, mild cognitive impairment, or dementia. Neuroepidemiology 2010;34:34-42 doi: 10.1159/000255464.

[19] R. Yaari, A. Fleisher, A. Gamst, V. Bagwell, L. Thal. Utility of the telephone interview for cognitive status for enrollment in clinical trials. Alzheimer’s & Dementia 2006;2:104-109 doi: 10.1016/j.jalz.2006.02.004.

[20] K. Fraser, J. Meltzer, F. Rudzicz, Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer's Disease 2016;49:407-422 doi: 10.3233/JAD-150520.

[21] T. Landauer, S. Dumais, A solution to Plato’s problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychological Review 1997;104:211-240 doi: https://doi.org/10.1037/0033-295X.104.2.211.

[22] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv preprint arXiv2013;1301.

[23] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations. NAACL-HLT 2018.

[24] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT 2018.

[25] T. Holmlund, J. Cheng, P. Foltz, *et al.* Updating verbal fluency analysis for the 21st century: Applications for psychiatry. Psychiatry Research 2019;273:767-769 doi 10.1016/j.psychres.2019.02.014.

[26] M. Rosenstein, P. Foltz, A. Vaskinn, B. Elvevåg, Practical issues in developing semantic frameworks for the analysis of verbal fluency data: A Norwegian data case study. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2015;124-133 doi 10.3115/v1/W15-1215.

[27] S. Pakhomov, L. Hemmy, A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. Cortex 2014;55:97-106 doi: 10.1016/j.cortex.2013.05.009.

[28] S. Pakhomov, L. Eberly, D. Knopman, Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. Neuropsychologia 2016;89:42-56 doi: 10.1016/j.neuropsychologia.2016.05.031.

[29] A.K. Troyer, M. Moscovitch, G. Winocur, Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. Neuropsychology 1997;11:138-146 doi: 10.1037//0894-4105.11.1.138.

[30] K. Nicodemus, B. Elvevåg, P. Foltz, M. Rosenstein, M., C. Diaz-Asper, D. Weinberger, Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. Cortex 2014;55:182-191 doi: 10.1016/j.cortex.2013.12.004.

[31] A. Schakel, B. Wilson, Measuring word significance using distributed representations of words. arXiv 2012;1508.02297v1.

[32] T. Holmlund, C. Chandler, P. Foltz, A. Cohen, J. Cheng, J. Bernstein, E. Rosenfeld, B. Elvevåg, Applying speech technologies to assess verbal memory. npj Digital Medicine 2020;3:33 doi: https://doi.org/10.1038/s41746-020-0241-7.

[33] C. Chandler, P. Foltz, J. Cheng, *et al.* Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership, in K. Niederhoffer, K. Hollingshead, P. Resnik, *et al* Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. 2019;137-147.

[34] W. Xu, J. Portanova, A. Chander, D. Ben-Zeev, T. Cohen, The centroid cannot hold: comparing sequential and global estimates of coherence as indicators of formal thought disorder. PsyArXiv 2020 doi: https://doi.org/10.31234/osf.io/sfkqc.

[35] K. Chapman, H. Bing-Canar, M. Alosco, E. Steinberg, B. Martin, C. Chaisson, *et al*. Mini Mental State Examination and Logical Memory scores for entry into Alzheimer's disease trials. Alzheimer's Research & Therapy 2016;8:9 doi: 10.1186/s13195-016-0176-z.

[36] L. Boise, M. Neal, J. Kaye, Dementia assessment in primary care: results from a study in three managed care systems. The Journals of Gerontology: Series A: Biological Sciences and Medical Sciences 2004;59A:621-626 doi: 10.1093/gerona/59.6.m621.

[37] S. Klekociuk, J. Summers, C. Vickers, M. Summers, Reducing false positive diagnoses in mild cognitive impairment: the importance of comprehensive neuropsychological assessment. European Journal of Neurology 2014;21:1330-1336 doi: 10.1111/ene.12488.

[38] Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behavioral and Brain Sciences. 2010;33(2-3):61-83. doi:10.1017/S0140525X0999152X

[39] K. Hitczenko, H.R. Cowan, M. Goldrick, V.A. Mittal. Racial and ethnic biases in computational approaches to psychopathology, Schizophrenia Bulletin 2021; sbab131, https://doi.org/10.1093/schbul/sbab131

[40] S. Blodgett, S. Barocas, H. Daume, H. Wallach. Language (technology) is power: a critical survey of “bias” in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 2020; 5454-5476. 10.18653/v1/2020.acl-main.485.

[41] C. Chandler, P.W. Foltz, B. Elvevåg. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. Schizophrenia Bulletin 2020; 46(1): 11–14, https://doi.org/10.1093/schbul/sbz105

[42] P. Foltz, M. Rosenstein, B. Elvevåg. Detecting clinically significant events through automated language analysis: Quo imus? npj Schizophr 2016; **2:** 15054 https://doi.org/10.1038/npjschz.2015.54

[43] A. Konig, A. Satt, A. Sorin, R. Hoory, A. Derreumaux, R. David, P.H. Robert. Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. Curr Alzheimer Res. 2018;15(2):120-129. doi: 10.2174/1567205014666170829111942.

Table 1: Demographic & descriptive characteristics of the sample (N=91).

Variables	Full Sample (N=91)	Cognitively Healthy (N=29)	aMCI (N=32)	AD (N=30)	*p*-value
Age (years)					
Mean ± SD	73.67 ± 6.94	72.48 ± 1.47	74.03 ± 1.01	74.93 ± 1.40	.486
Range	57-93	57-89	65-86	64-93	
Gender (n / (%))					
Male	41 (45)	12 (41)	13 (41)	15 (52)	.631
Female	50 (55)	17 (59)	19 (59)	14 (48)	
Education (years)					
Mean ± SD	17.35 ± 2.01	18.00 ± 0.37 *	17.34 ± 0.30	16.68 ± 0.43 *	.039
Range †	12-20	13-20	14-20	12-20	
Race/Ethnicity (n / (%))					
Non-Hispanic White	81 (90)	25 (86)	29 (91)	27 (93)	.665
Non-Hispanic Black	8 (9)	3 (10)	3 (9)	2 (7)	
Asian American	1 (1)	1 (3)	0 (0)	0 (0)	
Words: Semantic Fluency					
Mean ± SD	16.33 ± 7.44	21.03 ± 6.64*	16.75 ± 6.63*	11.37 ± 5.92*	<.001
Range	0-36	13-35	6-36	0-24	
Words: Free Speech					
Mean ± SD	328.98 ±166.74	372.35 ± 207.72	336.72 ± 152.28	278.8 ± 123.65	.092

Range	44-1110	118-1110	151-804	44-542	
TICS-M score					
Mean ± SD	35.26 ± 6.36	39.44 ± 0.63 *	36.03 ± 0.69 *	30.32 ± 1.31 *	<.001
Range ‡	18-49	34-47	28-49	18-44	
MMSE score					
Mean ± SD	27.26 ± 3.04	29.56 ± 0.12 *	28.22 ± 0.30 *	23.75 ± 0.51 *	<.001
Range §	19-30	28-30	24-30	19-29	

† the equivalent of a high school education is 12 years.

‡ possible range of scores = 0-50

§ possible range of scores = 0-30

Table 2: Summary of significant features extracted by NLP across two speech tasks. F-values are reported in brackets (all *p*-values <.05).

FREE SPEECH					
	Healthy vs aMCI vs AD	**Healthy vs combined AD/aMCI**	**Healthy vs AD**	**Healthy vs aMCI**	**AD vs aMCI**
Lexeme features	*Raw count* of nouns (3.91); determiners (3.62); modals (2.92); and present participle verbs (2.58)	*Raw count* of determiners (4.47); modals (4.32); and nouns (2.83)	*Raw count* of determiners (6.95); nouns (6.77); modals (6.18); and present participle verbs (2.54)	*Frequency* of past participle verbs (2.97); all verb types (2.39); non third person singular verbs (1.44); and modals (1.30)	*Raw count* of nouns (5.69); present participle verbs (5.59); determiners (3.79); and modals (1.80)
Syntactic features	Mean distance of all dependencies between words	Mean distance of all dependencies between words	Mean distance of all dependencies between words	--	Mean distance of all dependencies between words

	in a sentence (3.25)	in a sentence (3.21)	in a sentence (5.72)		in a sentence (3.58)
Semantic features	Mean coherence (5.93)	Mean coherence (11.07)	Mean coherence (11.07)	--	Mean coherence (6.55)
SEMANTIC FLUENCY					
	Number of unique animals spoken (16.31)	Number of unique animals spoken (19.64)	Number of unique animals spoken (33.78)	Number of categories of animals (6.05)	Number of unique animals spoken (11.27)
	Number of categories of animals (14.29)	Number of categories of animals (18.40)	Number of categories of animals (31.07)	Number of unique animals spoken (5.76)	Number of categories of animals (8.17)
	Maximum number of animals spoken per category (4.81)	Maximum number of animals spoken per category (6.86)	Maximum number of animals spoken per category (8.44)	Maximum number of animals spoken per category (2.59)	Maximum number of animals spoken per category (2.72)

Table 3: Confusion matrices. First panel: the cognitively healthy, aMCI, AD classifier; Second panel: the cognitively healthy vs aMCI/AD classifier; Third panel: cognitively healthy vs. aMCI in the most *clinically-relevant* model; Fourth panel: cognitively healthy vs AD in the *most clinically-relevant AND accurate* model; Fifth panel: aMCI vs AD in the *most clinically relevant* model.

Cognitively healthy, aMCI, AD classifier

		PREDICTED		
		HEALTHY	*aMCI*	*AD*
TRUE	*HEALTHY*	22	4	3
	aMCI	7	21	4
	AD	12	5	13

Cognitively healthy vs aMCI/AD classifier

		PREDICTED	
		HEALTHY	*aMCI/AD*

TRUE	*HEALTHY*	18	11
	aMCI/AD	5	57

Cognitively healthy vs. aMCI in the most *clinically-relevant* model

		PREDICTED		
		HEALTHY	*aMCI*	*AD*
TRUE	*HEALTHY*	22	4	3
	aMCI	7	21	4
	AD	12	5	13

Cognitively healthy vs AD in the *most clinically-relevant AND accurate* model.

		PREDICTED	
		HEALTHY	*AD*
TRUE	*HEALTHY*	27	2
	AD	5	25

aMCI vs AD in the *most clinically relevant* model.

		PREDICTED	
		aMCI	*AD*
TRUE	*aMCI*	29	3
	AD	10	20

FIGURE 1

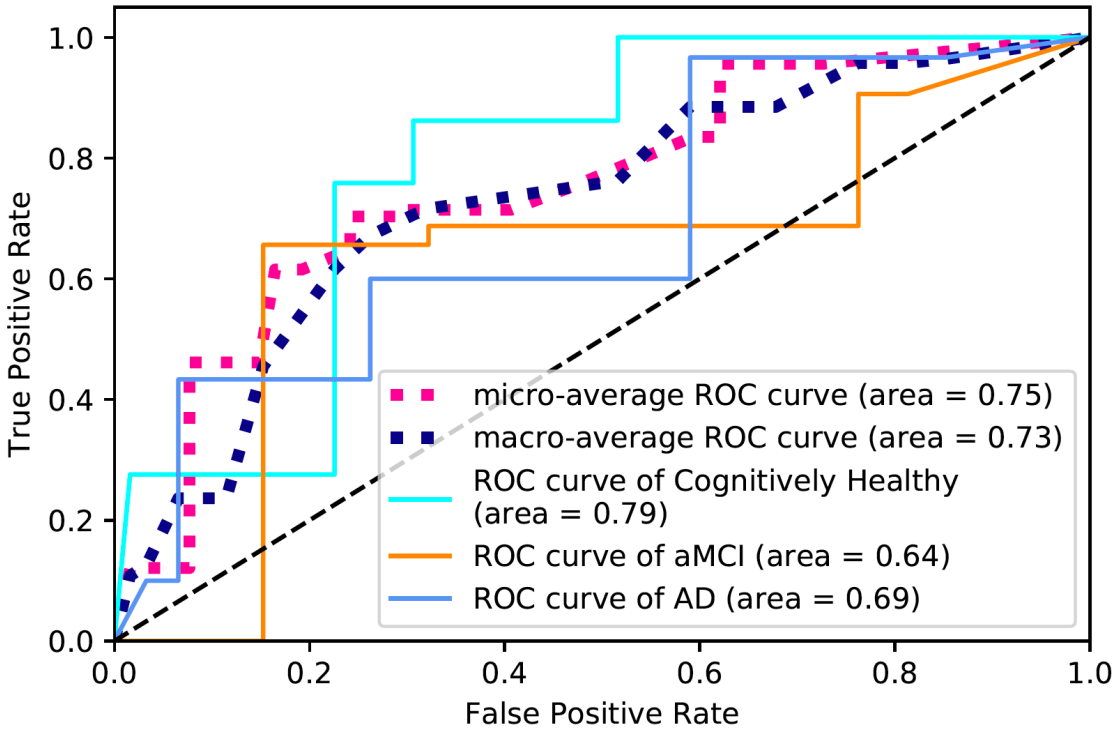

Figure 1: ROC curves of the three groups, as well as a micro- and macro-average. The model classified the cognitively healthy group with an AUC of 0.79, the aMCI group with an AUC of 0.64, and the AD group with an AUC of 0.69. It had an overall macro-average AUC of 0.73 and an overall micro-average AUC of 0.75. The macro-average computes the AUC independently for each group and then computes the average with each group treated equally, whereas the micro-average combines the contributions of all groups together.

FIGURE 2

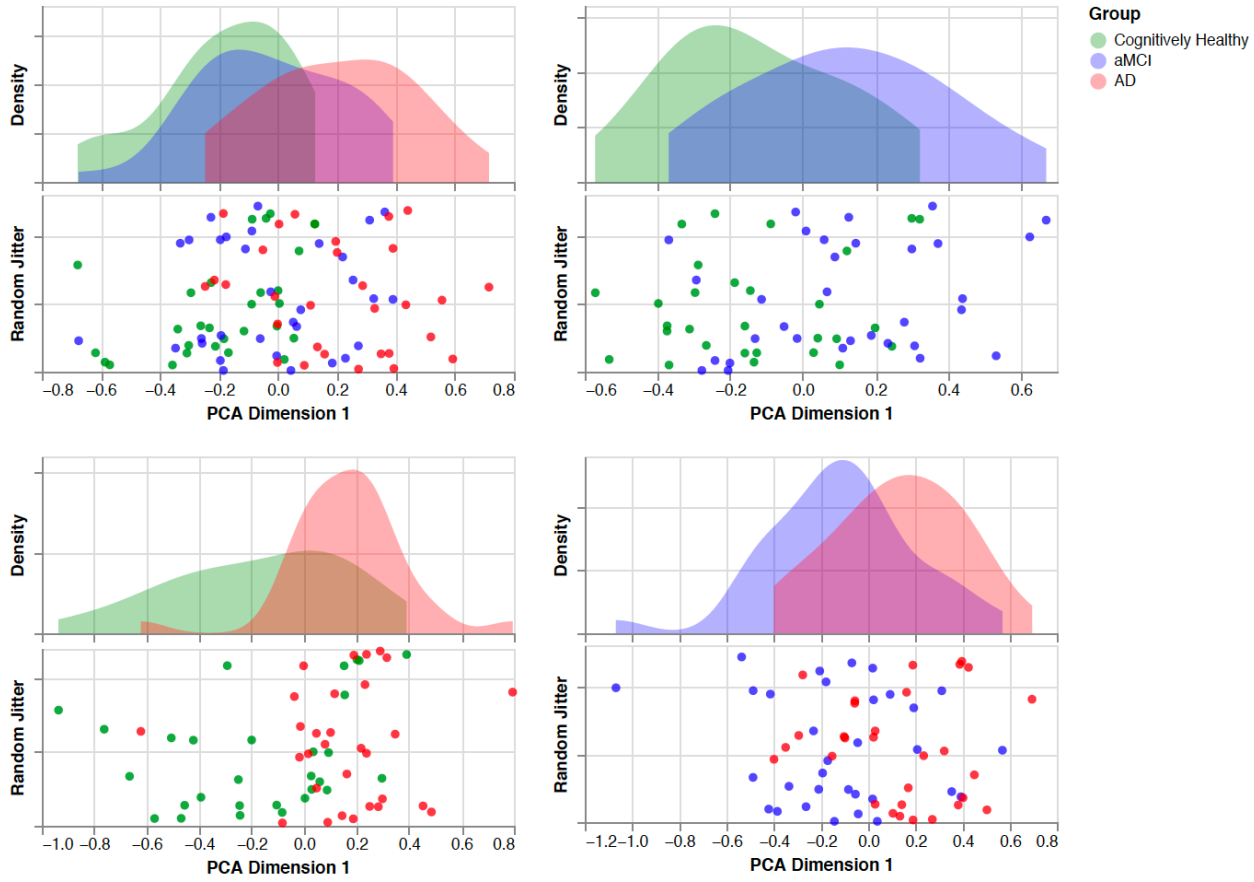

Figure 2: PCA dimensionality reduction of the top features used in the classification model showing the distributions for various experimental settings.

Top left panel: Using the top 3 features, applied to all three groups. The three groups are ordered as expected, with the peaks of cognitively healthy, aMCI, and AD ordered left to right with some overlap between each.

Top right panel: Using the top 6 features, applied to cognitively healthy vs aMCI. The two groups showed much overlap and tended to be difficult to differentiate from one another based on even their most discriminable features.

Bottom left panel: Using the top 6 features, applied to cognitively healthy vs AD. The two groups show some overlap, especially right at the peak of the AD group, but generally have distinct distributions with more discriminability.

Bottom right panel: Using the top 6 features, applied to aMCI vs AD. Again, the two groups show much overlap and tend to be difficult to differentiate from one another based on even their most discriminable features.

FIGURE 3

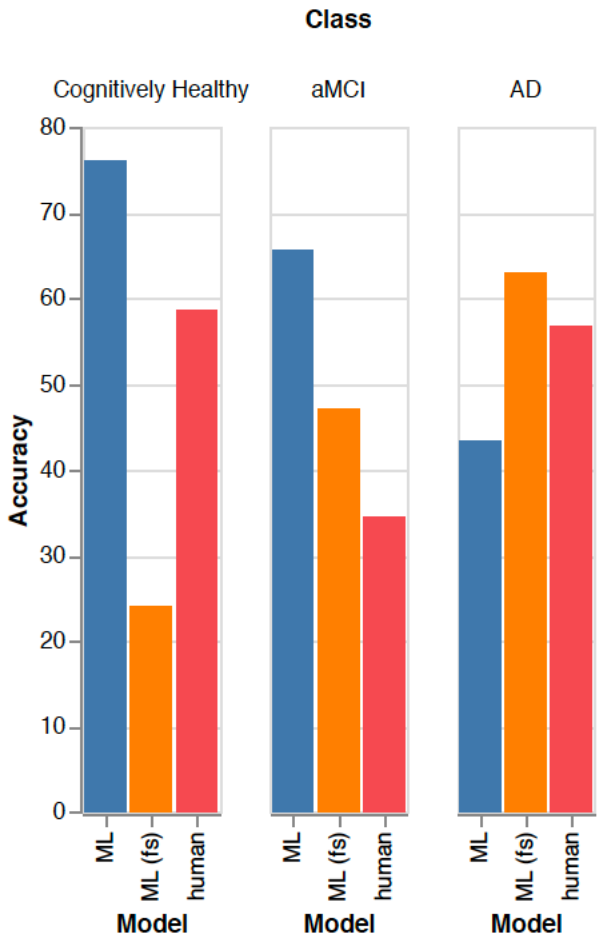

Figure 3: Accuracy of each model type (ML: the machine learning model based on the free speech and animal fluency tasks, ML (fs): the machine learning model based on the free speech task alone, and human) in each classification setting (Cognitively Healthy, aMCI, and AD).

APPENDIX A

Model Accuracy Metrics

Classification	Metric		
	Recall	*Precision*	*F1*
Cognitively Healthy, aMCI & AD			
Overall	0.62	0.62	0.62
Cognitively Healthy	0.76	0.54	0.63
aMCI	0.66	0.70	0.68
AD	0.43	0.65	0.52
Cognitively Healthy vs. Cognitive Decline			
Overall	0.87	0.87	0.87
Cognitively Healthy	0.69	0.87	0.77
Cognitive Decline	0.95	0.87	0.91
Cognitively Healthy vs. aMCI			
Overall	0.80	0.80	0.80
Cognitively Healthy	0.86	0.76	0.81
aMCI	0.75	0.86	0.80
Cognitively Healthy vs. AD			
Overall	0.88	0.88	0.88
Cognitively Healthy	0.93	0.84	0.88
AD	0.83	0.93	0.88
aMCI vs. AD			
Overall	0.70	0.79	0.79
aMCI	0.91	0.74	0.81
AD	0.67	0.87	0.75

APPENDIX B

Confusion matrices. Upper panel: the human diagnoses; Middle panel: the TICS-M; Lower panel: the MMSE.

Human diagnoses

		PREDICTED		
		HEALTHY	*aMCI*	*AD*
TRUE	*HEALTHY*	17	10	2
	aMCI	16	11	5
	AD	5	8	17

TICS-M

		PREDICTED		
		HEALTHY	*aMCI*	*AD*
TRUE	*HEALTHY*	29	0	0

	aMCI	30	2	0
	AD	12	8	10

MMSE

		PREDICTED		
		HEALTHY	*aMCI*	*AD*
TRUE	*HEALTHY*	29	0	0
	aMCI	25	6	1
	AD	4	10	15