Received: 16 May 2023 | Revised: 26 November 2023 | Accepted: 27 November 2023

DOI: 10.1002/dad2.12516

Alzheimer's & Dementia
Diagnosis, Assessment
& Disease Monitoring

RESEARCH ARTICLE

An explainable machine learning model of cognitive decline derived from speech

Chelsea Chandler[1] | Catherine Diaz-Asper[2] | Raymond S. Turner[3] | Brigid Reynolds[3] | Brita Elvevåg[4]

[1]Institute of Cognitive Science, University of Colorado, Boulder, Colorado, USA

[2]Department of Psychology, Marymount University, Arlington, Virginia, USA

[3]Department of Neurology, Georgetown University, Washington, District of Columbia, USA

[4]Department of Clinical Medicine, University of Tromsø – the Arctic University of Norway, Tromsø, Norway

Correspondence
Chelsea Chandler, Institute of Cognitive Science, University of Colorado Boulder, 1777 Exposition Drive, Boulder, CO 80309, USA.
Email: chelsea.chandler@colorado.edu

Funding information
National Institute on Aging, Grant/Award Number: R03AG052416

Abstract

INTRODUCTION: Traditional Alzheimer's disease (AD) and mild cognitive impairment (MCI) screening lacks the sensitivity and timeliness required to detect subtle indicators of cognitive decline. Multimodal artificial intelligence technologies using only speech data promise improved detection of neurodegenerative disorders.

METHODS: Speech collected over the telephone from 91 older participants who were cognitively healthy ($n = 29$) or had diagnoses of AD ($n = 30$) or amnestic MCI (aMCI; $n = 32$) was analyzed with multimodal natural language and speech processing methods. An explainable ensemble decision tree classifier for the multiclass prediction of cognitive decline was created.

RESULTS: This approach was 75% accurate overall—an improvement over traditional speech-based screening tools and a unimodal language-based model. We include a dashboard for the examination of the results, allowing for novel ways of interpreting such data.

DISCUSSION: This work provides a foundation for a meaningful change in medicine as clinical translation, scalability, and user friendliness were core to the methodologies.

KEYWORDS
Alzheimer's disease, cognitive screening, MCI, multimodal machine learning, NLP

Highlights

- Remote assessments and artificial intelligence (AI) models allow greater access to cognitive decline screening.
- Speech impairments differ significantly between mild AD, amnestic mild cognitive impairment (aMCI), and healthy controls.
- AI predictions of cognitive decline are more accurate than experts and standard tools.
- The AI model was 75% accurate in classifying mild AD, aMCI, and healthy controls.

1 | BACKGROUND

Early detection of cognitive decline in neurodegenerative disorders is urgently needed, with an estimated 153 million people worldwide affected by 2050.[1] Techniques requiring skilled medical professionals and well-resourced centers enable early detection (e.g., positron emission tomography [PET] scans, cerebrospinal fluid [CSF] markers, and structural magnetic resonance imaging [MRI]); however these methods are expensive, rarely available outside of large centers in wealthy countries, and require in-person attendance. Indeed, expanding access to diagnostics can ensure heightened equity in health care. A low-cost and remote diagnostic assay based on everyday speech collected over the telephone could provide a highly accessible screening method for neurodegenerative disorders.

Speech expression and content reflect the functioning of numerous cognitive processes including attention and memory.[2] Verbal communication involves various nuanced tasks: the formation of ideas, translation of thoughts into language, and articulation of utterances.[3] Advances in computational linguistics, acoustic processing, and machine learning (ML) enable the assessment of speech in a precise and reproducible manner[4–8] that may uncover new mechanistic knowledge regarding cognitive decline.[9–11] Loss of discourse complexity and connected speech are primary symptoms of cognitive decline associated with Alzheimer's disease (AD).[12–16] Decreased continuity of acoustic features, greater percentages of voiceless segments, and perturbations in speech amplitude are characteristic of patients diagnosed with AD.[17–20] There are also differences in features such as pause frequency, duration, and linguistic complexity between patients with mild cognitive impairment (MCI) and healthy controls.[21] More recently, the temporal integration of acoustics and language from spontaneous speech has been shown to be predictive of AD.[22]

Given the nuance of multimodal data, it was our goal to classify cognitive decline with transparency and explainability—an important aspect of AI in clinical settings.[23] Modalities are characterized by unique statistical properties, noise levels, and correlations to prediction variables, which necessitate consideration when combined in a model. It remains an open question which ML architectures can best represent nuanced multimodal human behavioral data.[24] One approach is to model all modalities together as one input, but this requires at least five training examples per feature dimension.[25] Alternatively, modalities can be accounted for with separate models and later combined with appropriate weightings.

This research aimed to pilot a remote and easily accessible telephone-based interview that requires no specialized equipment, is well-tolerated, and accessible at home,[26] and build a multimodal ML model for detecting early cognitive decline. Cognitively healthy individuals and those diagnosed with amnestic MCI (aMCI; MCI that primarily affects memory) or AD were administered short cognitive interviews, and multimodal features were extracted from responses (acoustic and language measurements from childhood memory recollections and animal fluency tasks). Improving on our earlier work,[9] we utilized language *and acoustic* processing methods and explainable ML models to create an accurate multiclass cognitive decline classifier. We explored various approaches for detecting cognitive decline with multimodal data and simple models. We designed a dashboard for researchers and clinicians to interpret the models and data, allowing new interpretations of experimental speech analysis in a detailed, yet understandable and meaningful manner.

RESEARCH IN CONTEXT

1. **Systematic review**: We reviewed traditional peer-reviewed articles and meeting abstracts. Although numerous studies have examined natural language processing for predicting cognitive decline, most have been conducted in highly controlled in-person environments, focusing mainly on language. Many studies lack machine learning model interpretability analyses or steps toward clinical implementation. Relevant citations are cited.
2. **Interpretation**: We demonstrate proof of principle that speech captured in naturalistic settings can be subjected to automated analyses and produce multimodal metrics that are more accurate and interpretable than expert humans and traditional dementia screening tests. The creation of a dashboard for interpreting model predictions promotes clinical translation.
3. **Future directions**: The current approach to collecting, analyzing, and interpreting data can potentially reach many more people at risk for cognitive decline than are currently assessed. However, model generalizability to demographically diverse populations and reliability of the models over time remain to be tested.

2 | METHODS

2.1 | Data

Participants comprised 91 older individuals who were cognitively unimpaired (n = 29, mean age = 72.48 [SD = 1.47], mean years of education = 18.00 [SD = 0.37], 41% female) or diagnosed with aMCI (n = 32, mean age = 74.03 [SD = 1.01], mean years of education = 17.34 [SD = 0.30], 41% female) or mild AD (n = 30, mean age = 74.93 [SD = 1.40], mean years of education = 16.68 [SD = 0.43], 52% female).[9,26] The group demographics did not differ significantly (p-values > 0.01). Clinical diagnoses followed established criteria separate from data collection. Participants completed a telephone interview, which included recalling a favorite childhood memory and naming as many animals as possible in 1 minute (animal fluency). The speech was audio-recorded and manually transcribed. The childhood memory task resulted in an average of 326.76 words spoken per participant (SD = 166.27, min = 44, max = 1110), and an average of 16.33 words (SD = 7.44, min = 0, max = 36) for the animal fluency task.

Participants also completed the Mini-Mental State Exam (MMSE[27]) and the modified Telephone Interview for Cognitive Status (TICS-M[28]). The three diagnostic groups differed significantly on both the MMSE and TICS-M, with the cognitively unimpaired participants performing the best (MMSE mean = 29.56 [SD = 0.12]; TICS-M mean = 39.44 [SD = 0.63]), the AD participants performing the worst (MMSE mean = 23.75 [SD = 0.51]; TICS-M mean = 30.32 [SD = 1.31]), and the aMCI participants performing intermediate between the other two (MMSE mean = 28.22 [SD = 0.30]; TICS-M mean = 36.03 [SD = 0.69]). The study was approved by Marymount University and Georgetown University institutional review boards (MU IRB #260).

2.2 | Feature extraction

Lexeme-level features were first extracted from free speech responses. Token and type counts measured poverty of speech. Part of speech counts and frequencies; type token ratio ($\frac{count(word\ types)}{count(word\ tokens)}$) (moving average type token ratio (*averaged moving windows of* ($\frac{count(word\ types)}{count(word\ tokens)}$); Brunét's index ($count(wordtokens)^{count(wordtypes)^{-0.165}}$); and content density $\frac{count(verbs+nouns+adjectives+adverbs)}{count(wordtokens)}$ measured poverty of content. Assays of verbigeration (uncontrollable word repetition) such as phrase- and word-level repetitions were counted. Counts of *um*'s, *ah*'s, and filler words derived indices of language fluency. Next, syntactic complexity features were computed, including statistics derived from sentence parse trees and speech graphs. Finally, semantic features were extracted. Measuring coherence, cosine distances between adjacent text windows (size ∈[2,8]; with LSA,[29] word2vec,[30] GloVe,[31] USE,[32] and BERT[33] embeddings) were calculated. Tangentiality was operationalized as the slope of cosine distances between the first and consecutive text windows, with the same parameters. With measuring of coherence, illogicality, and semantic paraphasia, a novel feature was created and implemented: statistics derived from BERT word probabilities in the context of a full response. Finally, sentence perplexities were extracted from BERT.

Traditionally, animal fluency is scored by counting the unique animals produced, ignoring significant structural and temporal information. Moving beyond this, objective quantifications of relationships between exemplars and categories of exemplars were explored. Troyer et al. proposed two metrics for measuring components of this task—clustering (producing words within one category, like *safari* or *pets*) and switching (changing between clusters).[34] This was implemented with hand-coded categories (supplied by the authors) and semantic word embedding distances. The semantic approach entails computing cosine distances of one animal's embedding to the next, and setting thresholds to determine whether the next animal belongs to a new category (when it falls below the threshold). The following is a segment of a response showing the Troyer categories (square brackets) and word embedding categories (BERT embeddings, threshold = 0.80; parentheses):

[(dog cat)] [(giraffes elephants lions tigers) (chimpanzees)]

Both approaches place *dog* and *cat* into one category (Pets; cosine distance > 0.8). The Troyer approach considers the subsequent animals all within the African category; however, the embedding approach splits *chimpanzees* into a new category as the cosine distance between *tigers* and *chimpanzees* falls below the threshold (0.52; the others having distances > 0.8). Finally, animal embedding magnitudes were computed, a measurement of typical word usage/familiarity (uncommon animals have larger magnitudes than common animals).

Acoustic features were extracted from both speech tasks. The Compare 2016 acoustic feature set[35]—including the Geneva Minimalistic Acoustic Parameter Set (GeMAPS[36])—was computed using openSMILE.[37] The Compare 2016 feature set aggregates frame-level acoustic properties over the entire file. Because there is reason to analyze acoustics at a frame-level (to measure how acoustics change with each other or with language), the files were segmented and acoustic features were computed for individual frames using a Pratt[38] script. Aggregate statistics, moving window averages, and feature correlations were computed over all frames.

Finally, cross-modal features were extracted from only the animal fluency task (due to poor automated word segmentation in free speech). Relationships between language and acoustics were computed to uncover interactions between these processes. One approach involved time-aligning modalities and correlating acoustic and language features. Another investigated frame acoustics for repeated versus non-repeated animals, and animals that begin new categories versus within-categories, motivated by exploring fluctuations in acoustics when important cognitive processes are occurring. We computed absolute differences of features binned by their cosine distance being greater than or less than a threshold. Moving averages and correlations were also computed.

2.3 | Multimodal modeling

Three approaches were explored for modeling multimodal, multitask data (Approaches A, B, C). We used Decision Tree Classifiers to retain simplicity and explainability. All models utilized *only* significant features (p values < 0.01) as computed by the F-statistic, a valuable univariate feature selection approach for small data sets that assesses the statistical significance of features in explaining class variability. This selection was carried out using the Python library sklearn, employing the feature_selection.SelectKBest and feature_selection.f_classif methods. Approach A involved creating multiple base models, each consisting of features from one modality and task (e.g., free speech language and acoustics, and animal fluency language, acoustics, and crossmodal), and using simple ensemble methods like *voting*, *most confidence*, and *summation* (Figure 1). *Voting* uses the majority vote of contributing model predictions. *Most confidence* uses the prediction of the base model with the highest probability (i.e., confidence). *Summation* sums the class probabilities among the base models, and uses the class with the largest summation. Approach B was similar, but with a meta-learner to learn weights for the submodels' predictions and generate the final prediction accordingly. Finaly, Approach C included all features in one input vector and used one model for prediction.

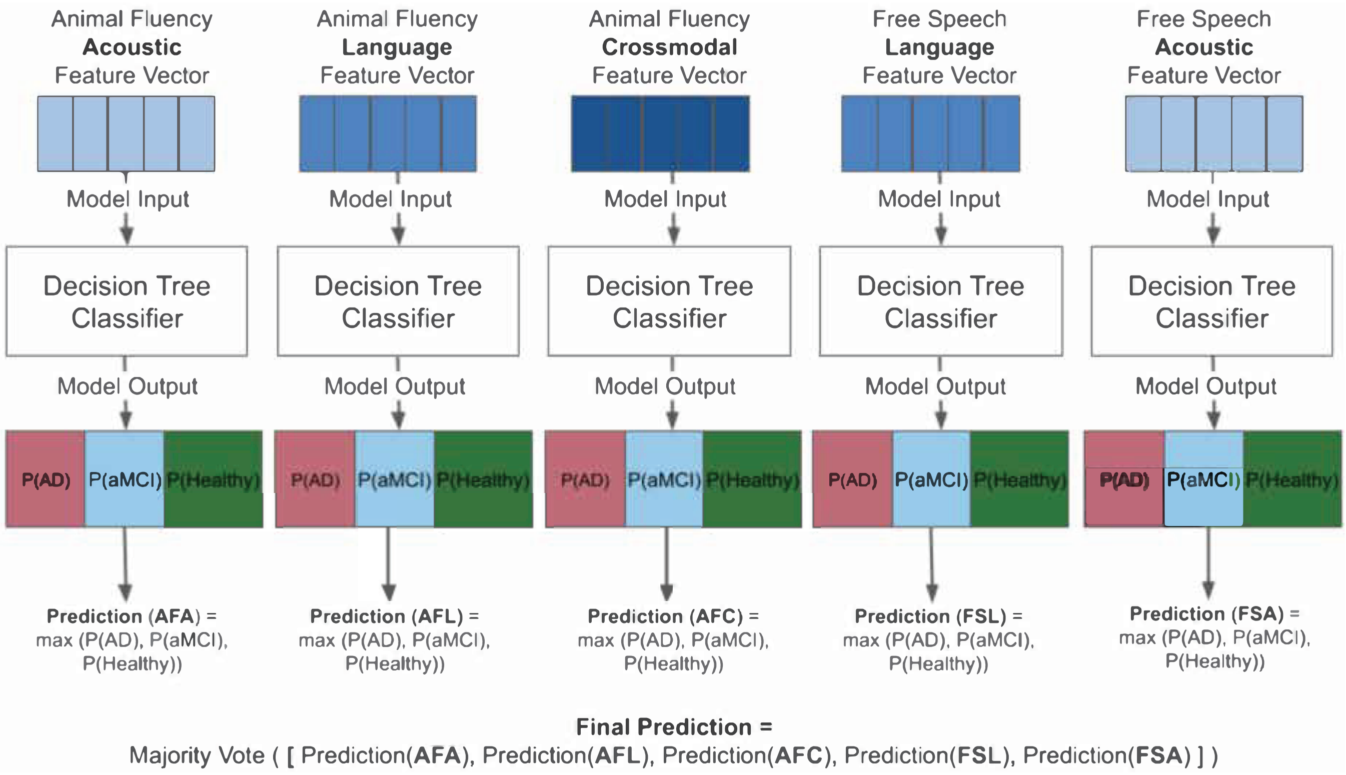
Animal Fluency
Acoustic
Feature Vector
Animal Fluency
Language
Feature Vector
Animal Fluency
Crossmodal
Feature Vector
Free Speech
Language
Feature Vector
Free Speech
Acoustic
Feature Vector
Model Input
Decision Tree Classifier
Model Output
P(AD)
P(aMCI)
P(Healthy)
Prediction (AFA) = max (P(AD), P(aMCI), P(Healthy))
Prediction (AFL) = max (P(AD), P(aMCI), P(Healthy))
Prediction (AFC) = max (P(AD), P(aMCI), P(Healthy))
Prediction (FSL) = max (P(AD), P(aMCI), P(Healthy))
Prediction (FSA) = max (P(AD), P(aMCI), P(Healthy))
Final Prediction =
Majority Vote ([Prediction(AFA), Prediction(AFL), Prediction(AFC), Prediction(FSL), Prediction(FSA)])

FIGURE 1 A visual representation of the ensemble model for cognitive decline. Five base models are created on unitask, unimodal data and the predictions from the five base models are combined with simple ensembling techniques.

3 | RESULTS

Highly predictive features for cognitive decline detection are provided. We examined whether acoustics would improve language-based cognitive decline models, so we present language-only models and explore approaches to creating explainable, multimodal models. Finally, we present an interface for interpreting the output.

3.1 | Animal fluency and free speech features

Language, acoustic, and crossmodal animal fluency features, and language and acoustic free speech features were predictive of diagnosis, with varied levels of discriminability. Table 1 lists the top features from each modality based on the F-statistic for differentiating classes. The highest-ranked animal fluency language features were variations of the maximum number of animals per category created with BERT embeddings. From free speech, the Mel Frequency Cepstral Coefficient (MFCC) acoustic features, BERT and USE coherence measures, and individual word probabilities from BERT were highly predictive. Principal component analysis (PCA) was used to reduce the set of most significant ($p < 0.01$) features from each modality per task for visualization purposes (Figure 2). From animal fluency, language separated healthy controls from cognitive decline in general, but was less suited to separate aMCI from AD. Acoustic differences were observed, with healthy participants distributing evenly over the peaks of the AD and aMCI classes. Crossmodal features differentiated the three classes well, with an expected ordering of healthy controls, aMCI, and AD. The free speech language modality differentiated the classes well with an expected ordering of healthy, aMCI, then AD. Acoustics show the three classes aligning, with a long tail extending out from the AD class distribution.

3.2 | Unimodal machine learning modeling

We previously presented five unimodal language-based prediction scenarios: (1) cognitively healthy versus aMCI versus AD, (2) cognitively healthy versus cognitive decline (aMCI and AD combined), (3) cognitively healthy versus aMCI, (4) cognitively healthy versus AD, and (5) aMCI versus AD.[3] The comparable scenario to the current study (Scenario 1) achieved 62% accuracy on our data set, compared to traditional screening tools TICS-M and MMSE (utilizing the best performing thresholds to differentiate groups[39–40] and scaling for education), which achieved 45% and 55% accuracy, respectively. Table 2 shows the confusion matrices for the ML model, TICS-M, and MMSE, with the ML approach showing an even spread of predictions compared to the overwhelming classification of healthy controls by TICS-M and MMSE.

TABLE 1 Top seven most predictive language, acoustic, and crossmodal features from the two tasks, sorted by F statistic.

Task	Language features	F, *p*-value	Acoustic Features	F, *p*-value	Crossmodal features	F, *p*-value
Animal fluency	Max animals per category (threshold = 0.8 and BERT)*	16.41, .000001	Correlation of *mean F0 slope octaves* & *max intensity*	10.77, .00006	*Max spectral slope* when cosine < .91 (BERT)*	7.34, .001
	Number animals spoken without repetitions	16.31, .000001	Correlation of *mean F0 slope* & *max intensity*	9.41, .0002	*Min F2* when cosine< .95 (BERT)*	6.80, .001
	Number categories (Troyer categories)*	14.29, .00001	Correlation of *F0 SD* & *max intensity*	6.97, .002	Absolute difference between *average alpha energy ratio* of cosines < .89 & > = .89 (BERT)	6.39, .002
	SD animals per category (threshold = 0.8 and BERT)*	12.38, .00001	Correlation of *max intensity* & *average MFCC1*	5.94, .003	Absolute difference between *average alpha density diff* of cosines < .89 & > = .89 (BERT)	6.31, .003
	Number categories / number animals (threshold = 0.85 and BERT)*	12.27, .00002	Max amplitude	5.86, .004	Absolute difference between *average MFCC2* of cosines < .89 & > = .89 (BERT)	6.31, .003
	Number categories (threshold = 0.8 and BERT)*	12.27,.00002	Min F1	4.97,.008	*SD harmonics to noise ratio* when cosine < .91 (BERT)*	5.87, .004
	Number animals (with repetitions)	9.17, .0004	Average amplitude	4.96, .009	Absolute difference between *average MFCC4* cosines < .95 & > = .95 (BERT)	5.17, .007
Free speech	Average coherence (USE embeddings, window size = 2)*	11.13, .00004	MFCC range	14.50, 0.000003		
	Minimum coherence (BERT embeddings, window size = 7)*	7.66, 0.0008	Percentage of time MFCC is above a 50% threshold	10.85, 0.00006		
	Proportion of words with BERT word probability < 0.001*	4.88, 0.009	Spectral slope left curvature time	9.28, 0.0002		
	SD of coherence (USE embeddings, window size = 3)*	4.33, 0.01	Spectral slope third quartile	8.09, 0.0005		
	Count of nouns	3.91, 0.01	MFCC standard deviation	7.98, 0.0006		
	BERT word probability moving average (window size = 2)	3.65, 0.01	MFCC maximum	7.92, 0.0006		
	Count of determiners	3.62, 0.01	Voicing probability of the final fundamental frequency candidate	7.82, 0.0007		

Acoustic features were smoothed over the entire response using moving average filters. If multiple feature variations were in the top features, the feature with the highest F-statistic is retained and subsequent variations are skipped to avoid redundancy (denoted with an asterisk).

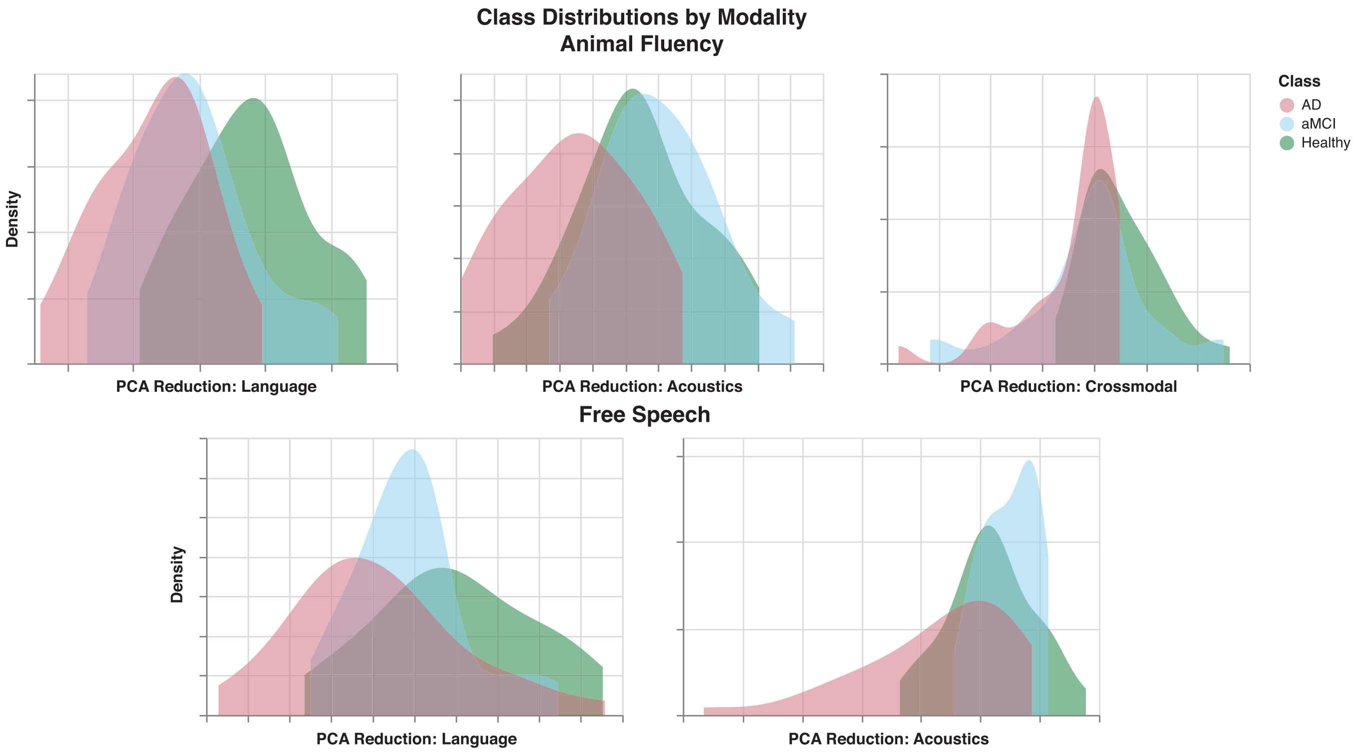

FIGURE 2 Density plots of the significant features (p values < 0.01) reduced with PCA reductions. Top row: language (left), acoustics (middle), and crossmodal (right) modalities from animal fluency. Bottom row: language (left) and acoustic (right) modalities from free speech. PCA, principal component analysis.

TABLE 2 Confusion matrix of the (1) unimodal machine learning-based classifier, (2) TICS-M test, and (3) MMSE test for classifications of cognitively healthy, aMCI, and AD participants.

		True		
		AD (n = 30)	aMCI (n = 32)	Healthy (n = 29)
Unimodal machine learning classifier				
Predicted	AD	**13**	4	3
	aMCI	5	**21**	4
	Healthy	12	7	**22**
TICS-M Test				
Predicted	AD	**10**	0	0
	aMCI	8	**2**	0
	Healthy	12	30	**29**
MMSE Test				
Predicted	AD	**15**	1	0
	aMCI	10	**6**	0
	Healthy	5	25	**29**

Abbreviations: AD, Alzheimer's disease; aMCI, amnestic Mild Cognitive Impairment; MMSE, Mini-Mental State Exam; TICS-M, Telephone Interview for Cognitive Status.

3.3 | Multimodal machine learning modeling

Five Decision Tree Classifiers were created using the most diagnostically salient features from language, acoustics, and crossmodal animal fluency modalities, and language and acoustics free speech modalities. Parameters were learned using leave-one-out cross-validation due to limited data set size. Animal fluency models achieved accuracies of 58% (language), 65% (acoustics), and 61% (cross-modal), whereas free speech models achieved accuracies of 56% (language) and 63% (acoustics). Precision, recall, and F1 scores for each model are

TABLE 3 (1) Precision, recall, and F1score broken down by participant class and macro averaged (computes the statistic individually per class and then averages the three together) and (2) confusion matrix for the best unimodal unitask ensemble model with voting (Approach A).

Classification metrics				
	Precision	**Recall**	**F1 score**	**Support**
Healthy	0.74	0.79	0.77	$n = 29$
aMCI	0.69	0.75	0.72	$n = 32$
AD	0.84	0.70	0.76	$n = 30$
Macro avg	0.76	0.75	0.75	$N = 91$
Confusion matrix				
		True		
		AD ($n = 30$)	**aMCI ($n = 32$)**	**Healthy ($n = 29$)**
Predicted	**AD**	**21**	3	1
	aMCI	6	**24**	5
	Healthy	3	5	**23**

supplied in Appendix A (Tables A1–A5). Results indicated that acoustics were more predictive than language. Different techniques for base model creation and their implications for accuracy are also detailed in Appendix A (e.g., creating the base models, aggregating by mode, then ensembling [Table A6]; and creating base models, aggregating by task, then ensembling [Table A7]).

Within each task, agreement between modality-specific predictions was assessed. Larger variations motivate an ensemble approach as each mode contributes different information. Had all models correlated highly with one another, limited gains would arise from ensembling (vs using one). For animal fluency, predictions for the healthy participants had moderate correlations across the three modes: language and acoustic (Pearson's $r = 0.28$, $p < 0.01$), language and crossmodal ($r = 0.39$, $p < 0.01$), acoustic and crossmodal ($r = 0.25$, $p = 0.01$). Correlations were higher within the AD class: language and acoustic ($r = 0.43$, $p < 0.01$), language and crossmodal ($r = 0.57$, $p < 0.01$), and acoustic and cross-modal ($r = 0.43$, $p < 0.01$). This was not the case for the aMCI class where there was no correlation in some cases: language and acoustic ($r = -0.02$, $p = 0.84$), language and crossmodal ($r = 0.09$, $p = 0.37$), acoustic and crossmodal ($r = 0.32$, $p < 0.01$). Higher correlations were expected with the crossmodal modality as it captures aspects of both acoustics and language. In free speech, correlations between acoustic and language modalities were lowest in the healthy class ($r = 0.25$, $p = 0.01$), higher in the aMCI class ($r = 0.35$, $p < 0.01$), and highest in the AD class ($r = 0.42$, $p < 0.01$). Model ensembling was motivated due to variance in predictions.

The highest accuracy model was both the voting ensemble (Approach A) and the Decision Tree Classifier meta-learner (Approach B), with 75% accuracy overall. However, the voting method was preferred as it avoids unnecessary overfitting to small data sets (Table 3).

Finally, one multimodal multitask cognitive decline model was explored (Approach C). ML models necessitate large data sets to learn the complexities of multidimensional data, and a sample size of 91 participants may be insufficient. Furthermore, the density plot of the multimodal, multitask data set reduced with PCA showed that the distribution resembled the animal fluency language modality and thus explained much of the variability in the data set, potentially overwhelming other modalities in one model. Feature importance extracted from the classifier showed that four of the top five features were from animal fluency language and one was from animal fluency acoustics. The accuracy of the classifier evaluated on the entire data set with leave-one-out cross-validation was 66%. Appendix A (Table A8) supplies precision, recall, and F1 scores.

Ensembling five unimodal unitask base models with voting generated the most accurate and explainable predictions. This approach was built from different data sources and harnessed the knowledge of multiple expert learners. Decision trees provided feature importances, and voting identified which modalities contributed to the prediction and how strongly they agreed. Model disagreement entails less algorithmic certainty, which is critical knowledge in clinical implementation as this may warrant human review. In terms of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV),[41] the voting classifier (AD sensitivity = 87.50%, specificity = 95.83%, PPV = 95.46%, NPV = 88.46%; aMCI sensitivity = 82.76%, specificity = 85.71%, PPV = 82.76%, NPV = 85.71%) had significantly more balanced outcomes overall than the TICS-M (AD sensitivity = 45.45%, specificity = 100%, PPV = 100%, NPV = 29.27%; aMCI sensitivity = 6.25%, specificity = 100%, PPV = 100%, NPV = 49.15%) and MMSE (AD sensitivity = 75%, specificity = 100%, PPV = 100%, NPV = 85.29%; aMCI sensitivity = 19.36%, specificity = 100%, PPV = 100%, NPV = 53.70%).

3.4 | Explainability

We created an interpretable user interface (Figure 3) for displaying model outputs, incorporating AI explainability, information visualization theory, and human-centered computing. The top section

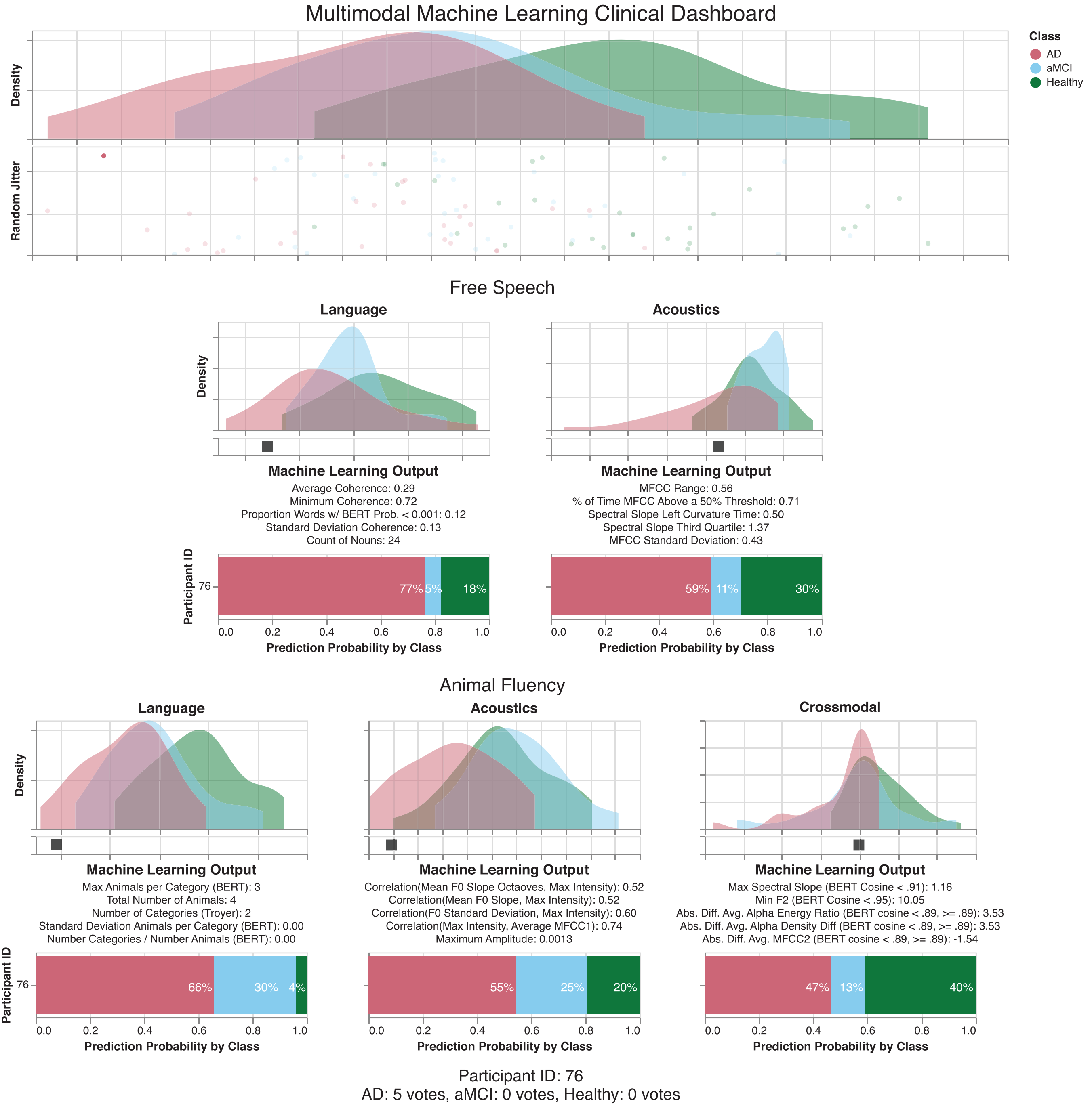

FIGURE 3 Machine learning interface displaying the selection of an individual with AD. The selected participant falls at the peak for free speech acoustics and animal fluency crossmodal (i.e., typical of those with AD), and at the extreme of the AD plot in the other modes (i.e., typical of the AD class and atypical of the others). The models were generally confident that the selected individual belonged to the AD class, with less confidence in the two modes where the participant was closer to the other classes. AD, Alzheimer's disease.

contains density and scatter plots of the first component of the full PCA-reduced data set. Hovering over points displays Participant's IDs. Selecting a point populates the interface with information from that individual. The selected point remains opaque, while others become transparent, allowing users to view an individual compared to the other participants and classes. Next are animal fluency and free speech sections with density plots and an indicator of where the selected participant falls compared to others within that task and modality. Values of the top features as determined by feature importance are shown, along with a stacked bar chart of the probabilities the model gave for each of the classes. Wider bars imply higher confidence. Explanations as multiclass probabilities were shown to be a superior technique for AI-based clinical decision-making support.[42] Finally, a breakdown of the model votes is given. Five votes for one class implies higher certainty than an even spread.

This interface harnesses several interaction categories from information visualization theory[43]: *Select* (clicking data in the scatterplot), *Reconfigure* (multiple view of data), *Elaborate* (showing more details, e.g., Participant ID, feature values, ML results), *Filter* (emphasizing selected individuals and filling graphs with corresponding data), and *Connect* (interactions with one plot affect others). This interface avoids the desert fog issue (zooming and panning causing the context to be unclear) by keeping non-selected data in scope to understand context. Colors were selected to be discernible to those with color blindness using the Coloring for Colorblindness tool.[44] A human-centered co-design process[45] brought together clinicians, neurocognitive assessment experts, and AI researchers, satisfying diverse expectations and constraints from clinical and AI perspectives.

4 | DISCUSSION

A speech-based AI model may be more sensitive than traditional cognitive screeners to early dementia detection, yet an estimation of cost efficacy is crucial to promote uptake in clinical settings. Such analyses are complicated, and the resulting economic value frameworks will be dependent upon relevant jurisdictions. Previous studies suggest that AI-based cognitive impairment diagnostic tools can be cost saving (e.g., estimated net monetary benefits per person to the UK National Health Service of £154/$193USD in primary care and £281/$352USD in memory clinic settings[46]). To assess the viability and impact of the proposed speech-based assessment tool, our future work underscores the need for a cost-effectiveness analysis, which is essential for informed decision-making and integration into clinical practice.

Speech-based biomarkers will contribute to the early detection of neurodegenerative diseases, complementing established diagnostic methods (e.g., blood-based, brain scans), with the added advantage of more accessible, widespread, and remote administrations. Early diagnosis confers important benefits to diagnosed individuals, caregivers, loved ones, and society. In addition to providing significant medical, emotional, and social benefits and facilitating participation in clinical trials, early diagnosis enables individuals to prepare legal, financial, and end-of-life plans while they are still cognitively able to make decisions and share their wishes.

In this research, clinical relevance of model features was a priority. To maximize language output, our interviewer elicited more speech in the event of short responses. Thus word count was not a highly predictive indicator of cognitive decline (as it often is[47,48]). Rather than considering this a confound, collecting speech in this manner allowed features like coherence to be indicators of cognitive decline without being impacted by word count.[49] For fluency tasks, neurodegenerative disorders can impair the amount, usualness, manner, and ordering of items retrieved.[50] Thus our animal fluency language features sought to measure these aspects of retrieval.

Acoustic findings aligned with literature, showing greater similarity between the healthy and aMCI classes than with AD. Healthy and aMCI classes had larger MFCC ranges and higher maximum amplitudes compared to the AD class, indicating less restriction in speech power. Correlations between F0 slope and maximum intensity were approximately zero for the healthy and aMCI classes, but positive for the AD class, revealing unique speech patterns within the AD class. Despite imperfect audio recordings, our research was successful in extracting meaningful signals (detailed in Appendix B). Although acoustic features can be influenced by demographics, medications, and medical conditions, research has shown that certain acoustic features may generalize beyond these factors and thus were retained as part of our framework.

This research was limited by a small and homogenous data set, which increases the risk of overfitting, given the inability to evaluate using a separate data set. Mitigation efforts included adopting simple model architectures that promote balanced and generalized predictions. For instance, harnessing a decision tree (as opposed to voting) as a meta-learner involves hyperparameter choices and learned weights, possibly resulting in overfitting and capturing irrelevant noise. Despite these limitations, we demonstrated improvements in constrained prediction scenarios through thoughtful multimodal feature aggregation. With a small data set, we found little difference between varied ML architectures. This is beneficial—showing that features were impervious to modeling choices and were themselves predictive independent of particular algorithms, and a weakness—as various architectures *are* more well suited for different data types; however, this variation could not be utilized. Further testing on larger, more diverse data sets is necessary to test for generalization.

AUTHOR CONTRIBUTIONS

Chelsea Chandler: Data Curation; Software; Methodology; Formal Analysis; Visualization; Writing-Original Draft; Reviewing and Editing. Catherine Diaz-Asper: Conceptualization; Methodology; Visualization; Writing-Original Draft; Reviewing and Editing; Project Administration; Funding Acquisition. Raymond S. Turner: Resources. Brigid Reynolds: Resources. Brita Elvevåg: Supervision; Writing-Original Draft; Reviewing and Editing.

ACKNOWLEDGMENTS

The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript. The authors would like to thank the study participants for their time. This work was supported by the National Institute on Aging (grant numberR03AG052416).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest. Author disclosures are available in the supporting information.

CONSENT STATEMENT

All human subjects provided informed consent.

ORCID

Chelsea Chandler https://orcid.org/0000-0002-9409-0937

REFERENCES

1. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050. *Lancet Public Health*. 2022. doi:10.1016/S2468-2667(21)00249-8
2. Roger E, Banjac S, Thiebaut de Schotten M, Baciu M. Missing links: the functional unification of language and memory (L∪M). *Neurosci Biobehav Rev*. 2022;133:104489. doi:10.1016/j.neubiorev.2021.12.012 PMID: 34929226
3. Dronkers N, Ogar J. Brain areas involved in speech production. *Brain*. 2004;127(7):1461-1462.
4. Chandler C, Foltz PW, Cohen AS, et al. Machine learning for ambulatory applications of neuropsychological testing. *Intell Based Med*. 2020;1-2:100006. doi:10.1016/j.ibmed.2020.100006
5. Holmlund TB, Chandler C, Foltz PW, et al. Applying speech technologies to assess verbal memory in patients with serious mental illness. *NPJ Digit Med*. 2020;3:33. doi:10.1038/s41746-020-0241-7
6. Holmlund TB, Fedechko TL, Elvevåg B, Cohen AS. Tracking language in real time in psychosis. A Clinical Introduction to Psychosis: Foundations For Clinical Psychologists and Neuropsychologists *663-685*. Elsevier Academic Press; 2020. doi:10.1016/B978-0-12-815012-2.00028-6
7. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngoscope Investig Otolaryngol*. 2020;5(1):96-116. doi:10.1002/lio2.354
8. Voleti R, Liss JM, Berisha V. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE Trans Audio Speech Lang Process*. 2020;14(2):282-298. doi:10.1109/jstsp.2019.2952087
9. Diaz-Asper C, Chandler C, Turner RS, Reynolds B, Elvevåg B. Increasing access to cognitive screening in the elderly: applying natural language processing methods to speech collected over the telephone. *Cortex*. 2022;156:26-38. PMID: 36179481. doi:10.1016/j.cortex.2022.08.005
10. Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*. 2020:28.
11. Orimaye S, Wong J, Wong C. Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. *PLoS ONE*. 2018;13:e0205636.
12. Berisha V, Wang S, LaCross A, Liss J. Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case study comparing the press conferences of Presidents Ronald Reagan and George Herbert Walker Bush. *J Alzheimers Dis*. 2015;45(3):959-963. doi:10.3233/JAD-142763
13. Filiou RP, Bier N, Slegers A, Houzé B, Belchior P, Brambati SM. Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: a scoping review. *Aphasiology*. 2020;34(6):723-755. doi:10.1080/02687038.2019.1608502
14. Snowdon DA, Kemper SJ, Mortimer JA, Greiner LH, Wekstein DR, Markesbery WR. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the Nun Study. *JAMA*. 1996;275(7):528-532.
15. Pakhomov S, Chacon D, Wicklund M, Gundel J. Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. *Behav Res Methods*. 2011;43(1):136-144. doi:10.3758/s13428-010-0037-9
16. Mueller KD, Van Hulle CA, Koscik RL, et al. Amyloid beta associations with connected speech in cognitively unimpaired adults. *Alzheimer's Dement*. 2021;13:e12203. doi:10.1002/dad2.12203
17. Shimoda A, Yue L, Hayashi H, Kondo N. Dementia risks identified by vocal features via telephone conversations: a novel machine learning prediction model. *PLoS ONE*. 2021;16(7):e0253988. doi:10.1371/journal.pone.0253988
18. König A, Satt A, Sorin A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement*. 2015;1(1):112-124. doi:10.1016/j.dadm.2014.11.012
19. Meilán JJG, Martínez-Sánchez F, Carro J, López DE, Millian-Morell L, Arana JM. Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? *Dement Geriatr Cogn Disord*. 2014;37(5-6):327-334. doi:10.1159/000356726
20. Meilán JJG, Martínez-Sánchez F, Carro J, Sánchez JA, Pérez E. Acoustic markers associated with impairment in language processing in Alzheimer's disease. *Span J Psychol*. 2012;15(2):487-494. doi:10.5209/rev_SJOP.2012.v15.n2.38859
21. Roark B, Mitchell M, Hosom JP, Hollingshead K, Kaye J. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Trans Audio Speech Lang Process*. 2011;19(7):2081-2090. doi:10.1109/TASL.2011.2112351
22. Martinc M, Haider F, Pollak S, Luz S. Temporal integration of text transcripts and acoustic features for Alzheimer's diagnosis based on spontaneous speech. *Front Aging Neurosci*. 2021;13:642647. doi:10.3389/fnagi.2021.642647
23. Chandler C, Foltz PW, Elvevåg B. Using machine learning in Psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophr Bull*. 2020;46(1):1114. doi:10.1093/schbul/sbz105
24. Cohen AS, Cox CR, Masucci MD, et al. Digital phenotyping using multimodal data. *Curr Behav Neurosci Rep*. 2020;7(4):212-220. doi:10.1007/s40473-020-00215-4
25. Koutroumbas K, Theodoridis S. *Pattern Recognition* 4th ed. Academic Press; 2008.
26. Diaz-Asper C, Chandler C, Turner RS, Reynolds B, Elvevåg B. Acceptability of collecting speech samples from the elderly via the telephone. *Digit Health*. 2021. doi:10.1177/20552076211002103 PMID: 33953936
27. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*. 1975;12(3):189-198. doi:10.1016/0022-3956(75)90026-6
28. Brandt J, Spencer M, Folstein M. The telephone interview for cognitive status. *Neuropsychiatry Neuropsychol Behav Neurol*. 1988;1(2):111-117.
29. Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol Rev*. 1997;104(2):211-240. doi:10.1037/0033-295X.104.2.211
30. Mikolov T, Chen K, Corrado G, Dean J, Efficient Estimation of Word Representations in Vector Space. *ArXiv*:1301.3781 [Cs]. 2013. doi:arxiv.org/abs/1301.3781
31. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics. 2014: 1532-1543.
32. Cer D, Yang Y, Kong S et al. Universal sentence encoder for English. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2018;169-174. doi:10.18653/v1/D18-2029
33. Devlin J, Chang MW, Lee K, Toutanova KB. Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1(Long and Short Papers). 2019:4171-4186. doi:10.18653/v1/N19-1423
34. Troyer AK, Moscovitch M, Winocur G. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology*. 1997;11(1):138-146. doi:10.1037//0894-4105.11.1.138
35. Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity & native lan-

guage. *Interspeech*. 2016. doi:10.21437/Interspeech.2016-129. *2016*, 2001-2005.

36. Eyben F, Scherer KR, Schuller BW, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput*. 2016;7(2):190-202. doi:10.1109/TAFFC.2015.2457417
37. Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*. 2010. 1459-1462. doi:10.1145/1873951.1874246
38. Boersma P, Weenink D. Praat: Doing Phonetics by Computer. 2022.
39. Knopman D, Roberts RO, Geda YE, et al. Validation of the telephone interview for cognitive status-modified in subjects with normal cognition, mild cognitive impairment, or dementia. *Neuroepidemiology*. 2010;34:34e42. doi:10.1159/000255464
40. Chapman K, Bing-Canar H, Alosco ML, et al. Mini mental state examination and logical memory scores for entry into Alzheimer's disease trials. *Alzheimers Res Ther*. 2016;8:9. doi:10.1186/s13195-016-0176-z
41. Sensitivity TR. Specificity, and Predictive values: foundations, pliabilities, and pitfalls in research and practice. *Front Public Health*. 2017;5:307. PMID: 29209603; PMCID: PMC5701930. doi:10.3389/fpubh.2017.00307
42. Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. *Nat Med*. 2020;26:1229-1234. doi:10.1038/s41591-020-0942-0
43. Yi JS, Kang YA, Stasko J, Jacko JA. Toward a deeper understanding of the role of interaction in information visualization. In IEEE Transactions on Visualization and Computer Graphics. 2007;13(6):1224-1231. doi:10.1109/TVCG.2007.70515
44. Nichols D. Coloring for Colorblindness. Accessed September 10, 2021. https://davidmathlogic.com/colorblind
45. Penuel WR, Roschelle J, Shechtman N. Designing formative assessment software with teachers: an analysis of the co-design process. *Res Pract Technol Enhanc Learn*. 2007;2(1):51-74.
46. Shore J, Kalafatis C, Stainthorpe A, Modarres MH, Khaligh-Razavi SM. Health economic analysis of the integrated cognitive assessment tool to aid dementia diagnosis in the United Kingdom. *Front Public Health*. 2023;11:1240901. doi:10.3389/fpubh.2023.1240901
47. Ostrand R, Gunstad J. Using automatic assessment of speech production to predict current and future cognitive function in older adults. *J Geriatr Psychiatry Neurol*;34(5):357-369. doi:10.1177/0891988720933358
48. Shao Z, Janse E, Visser K, Meyer AS. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Front Psychol*. 2014:5.
49. Hitczenko K, Cowan H, Mittal V, Goldrick M. Automated coherence measures fail to index thought disorder in individuals at risk for psychosis. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. 129-150. Online. Association for Computational Linguistics. 2021.
50. Henry JD, Crawford JR, Phillips LH. Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia*. 2004;42(9):1212-1222. doi:10.1016/j.neuropsychologia.2004.02.001 PMID: 15178173
51. Becker JT, Boller F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch Neurol*. 1994;51(6):585-594.
52. Green Forge Coop. MOSQITO [Computer software]. https://doi.org/10.5281/zenodo.5284054

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Chandler C, Diaz-Asper C, Turner RS, Reynolds B, Elvevåg B. An explainable machine learning model of cognitive decline derived from speech. *Alzheimer's Dement*. 2023;15:e12516. https://doi.org/10.1002/dad2.12516

APPENDIX A: DETAILED MULTIMODAL MODELING RESULTS

The animal fluency language model was 58% accurate overall. Precision, recall, and F1-score by class are given in Table A1.

TABLE A1 Precision, recall, and F1 score broken down by participant class and macro averaged (computes the statistic individually per class and then averages the three together) for the animal fluency language base model.

	Precision	Recall	F1 score	Support
Healthy	0.61	0.76	0.68	*N* = 29
aMCI	0.51	0.56	0.54	*N* = 32
AD	0.63	0.41	0.50	*N* = 29*
Macro avg	0.59	0.58	0.57	*N* = 90

Abbreviations: AD, Alzheimer's Disease; aMCI, amnestic Mild Cognitive Impairment.

The animal fluency acoustic model was 65% accurate overall. Precision, recall, and F1-score by class are given in Table A2.

TABLE A2 Precision, recall, and F1 score broken down by participant class and macro averaged (computes the statistic individually per class and then averages the three together) for the animal fluency acoustic base model.

	Precision	Recall	F1 score	Support
Healthy	0.68	0.68	0.68	*N* = 29
aMCI	0.55	0.66	0.60	*N* = 32
AD	0.78	0.62	0.69	*N* = 29*
Macro avg	0.67	0.65	0.66	*N* = 90

Abbreviations: AD, Alzheimer's Disease; aMCI, amnestic Mild Cognitive Impairment.

The animal fluency crossmodal model was 61% accurate overall. Precision, recall, and F1 score by class are given in Table A3.

TABLE A3 Precision, recall, and F1 score broken down by participant class and macro averaged (computes the statistic individually per class and then averages the three together) for the animal fluency crossmodal base model.

	Precision	Recall	F1-score	Support
Healthy	0.61	0.61	0.61	*N* = 29
aMCI	0.53	0.59	0.56	*N* = 32
AD	0.72	0.62	0.67	*N* = 29*
Macro avg	0.62	0.61	0.61	*N* = 90

Abbreviations: AD, Alzheimer's Disease; aMCI, amnestic Mild Cognitive Impairment.

* A participant from the AD class did not produce any animals in the animal fluency task and thus was excluded from this portion of the analyses.

The free speech language model was 56% accurate overall. Precision, recall, and F1 score by class are given in Table A4.

TABLE A4 Precision, recall, and F1 score broken down by participant class and macro averaged (computes the statistic individually per class and then averages the three together) for the free speech language base model.

	Precision	Recall	F1 score	Support
Healthy	0.62	0.69	0.66	*N* = 29
aMCI	0.50	0.44	0.47	*N* = 32
AD	0.55	0.57	0.56	*N* = 30
Macro avg	0.56	0.56	0.56	*N* = 91

Abbreviations: AD, Alzheimer's Disease; aMCI, amnestic Mild Cognitive Impairment.

The free speech acoustic model was 63% accurate overall. Precision, recall, and F1 score by class are given in Table A5.

TABLE A5 Precision, recall, and F1 score broken down by participant class and macro averaged (computes the statistic individually per class and then averages the three together) for the free speech acoustic base model.

	Precision	Recall	F1 score	Support
Healthy	0.53	0.55	0.54	*N* = 29
aMCI	0.68	0.72	0.70	*N* = 32
AD	0.67	0.60	0.63	*N* = 30
Macro avg	0.63	0.62	0.62	*N* = 91

Abbreviations: AD, Alzheimer's Disease; aMCI, amnestic Mild Cognitive Impairment.

Details on the multimodal unitask split of the base models are given here. This technique is first explored with a three-tiered ensemble approach. First, a multimodal animal fluency ensemble model and a multimodal free-speech ensemble model were created by harnessing the aforementioned base models. The best animal fluency model was 66% accurate overall using the voting mechanism with the three unimodal base models, and the best free-speech model was 65% accurate overall using a scaled (by overall class accuracy) summation mechanism. Alternatively, one could build one multimodal base model for the animal fluency task and one multimodal base model for the free speech task (without first learning within modalities) and combine them. Similar results occurred in these variations of the unitask models. The animal fluency model was the same at 66% accurate overall and the free-speech model was 66% overall. The best final prediction model based on two multimodal unitask models was 66% accurate overall ensemble with the summation technique (with no additional scaling for accuracy). Surprisingly, there was no boost in accuracy when combining models in this manner. Precision, recall, and F1 score by class for each model are supplied in Table A6.

The multimodal unitask ensemble model was 66% overall. Precision, recall, and F1-score by class are given in Table A6.

TABLE A6 Precision, recall, and F1 score broken down by participant class and macro averaged (computes the statistic individually per class and then averages the three together) for the multimodal unitask ensemble model.

	Precision	Recall	F1 score	Support
Healthy	0.59	0.68	0.63	*N* = 29
aMCI	0.64	0.66	0.65	*N* = 32
AD	0.79	0.66	0.72	*N* = 30
Macro avg	0.67	0.66	0.67	*N* = 91

Abbreviations: AD, Alzheimer's Disease; aMCI, amnestic Mild Cognitive Impairment.

In building a unimodal multitask ensemble model, all features from each modality, regardless of the task they were extracted from were first combined into a single model and the three base models were ensembled. As in the last scenario, two approaches can be taken. The first approach is by harnessing the unimodal unitask models to first create the base ensemble models. The language model ensembled the animal fluency language model and the free-speech language model, the acoustic model ensembled the animal fluency acoustic model and the free-speech acoustic model, and the crossmodal model remains the same regardless of approach as this modality only exists for the animal fluency data. Ensembling the three models created in this manner results in a best overall accuracy of 65% using summation (both scaled by overall accuracy and non-scaled) and voting. The second approach is to build three base models directly from the features: a language model that combines all language features from both tasks, an acoustic model that combines all acoustic features from both tasks, and finally the crossmodal model. Ensembling the three models created in this manner results in a best overall accuracy of 65% again using summation (both scaled by overall accuracy and non-scaled) and voting. As this approach could be deemed as more explainable, since there is not an obscuring with an additional round of ensembling, only this model's results are supplied in Table A7.

The unimodal multitask ensemble model was 65% overall. Precision, recall, and F1nscore by class are given in Table A7.

TABLE A7 Precision, recall, and F1 score broken down by participant class and macro averaged (computes the statistic individually per class and then averages the three together) for the unimodal multitask ensemble model.

	Precision	Recall	F1 score	Support
Healthy	0.68	0.61	0.64	*N* = 29
aMCI	0.63	0.65	0.64	*N* = 32
AD	0.65	0.69	0.67	*N* = 30
Macro avg	0.65	0.65	0.65	*N* = 91

Abbreviations: AD, Alzheimer's Disease; aMCI, amnestic Mild Cognitive Impairment.

The multimodal model with no special architectural considerations was 66% overall. Precision recall, and F1 score by class are given in Table A8.

TABLE A8 Precision, recall, and F1 score broken down by participant class and macro averaged (computes the statistic individually per class and then averages the three together) for the multimodal model with no special architectural considerations.

	Precision	Recall	F1 score	Support
Healthy	0.61	0.76	0.68	*N* = 29
aMCI	0.61	0.53	0.47	*N* = 32
AD	0.70	0.63	0.67	*N* = 30
Macro avg	0.65	0.65	0.65	*N* = 91

Abbreviations: AD, Alzheimer's Disease; aMCI, amnestic Mild Cognitive Impairment.

APPENDIX B: ACOUSTIC QUALITY

An impressive aspect of this work is that predictive acoustic information could be extracted given the recording quality. As many dementia studies harness the DementiaBank speech data set [B1], we regarded this as baseline audio quality and sought to determine how the current data set—collected out of the laboratory and over the telephone—compared. We compared our recordings to a demographically matched subset of the relatively older DementiaBank data set using the MOSQITO sound quality software [B2]. The DementiaBank cookie theft task was matched to our free speech task: 30 individuals with AD, 11 with MCI, and 28 healthy individuals from DementiaBank were age-matched within 4 years (with three exceptions from 89+-year-old participants) and on gender and ethnicity. The DementiaBank animal fluency task was matched in the same manner resulting in 28 individuals with AD, 12 with MCI, and 3 healthy individuals. For the free speech comparison, random 10-second segments of fluid speech were extracted from each file. All animal fluency audio was analyzed in both data sets. Average Tone-to-Noise Ratios and Prominence Ratios were computed for each sample and the average, minimum, maximum, and variance of these measures were compared. These metrics measure the level of tone relative to background sound level to determine how prominent the tone is to listeners. It has been shown that for Tone-to-Noise Ratio, the tone must be at least 8 dB above the level of noise to be audible, and discrete tones are said to be prominent if it is greater than or equal to 9 dB. Results of the audio quality analysis are given in Table B1.[51,52] In general, DementiaBank recordings had higher averages, minimums, maximums, and standard deviations than our data, implying that the audio quality of our data set was lower than that of a commonly analyzed data set in dementia research. Nevertheless, we found that meaningful acoustic features could be extracted from imperfect recordings, implying that elderly people can harness easily-available and low-cost equipment from home to generate sensitive predictions of cognitive decline rather than needing to visit a facility for assessment.

TABLE B1 Comparison of the audio quality of the current dataset with the DementiaBank data set.

Free speech (current data set)				Cookie theft (DementiaBank data set)			
Average tone-to-noise ratio	Average prominence ratio			Average tone-to-noise ratio		Average prominence ratio	
Avg:	10.30 dB	Avg:	12.83 dB	Avg:	11.19 dB*	Avg:	15.01 dB*
Min:	8.38 dB	Min:	9.21 dB	Min:	10.42 dB*	Min:	10.31 dB*
Max:	12.71 dB*	Max:	20.54 dB	Max:	12.00 dB	Max:	21.02 dB*
SD:	3.47 dB*	SD:	3.01 dB	SD:	0.63 dB	SD:	3.47 dB*
Animal fluency (current data set)				Animal fluency (DementiaBank data set)			
Average tone-to-noise ratio	Average prominence ratio			Average tone-to-noise ratio		Average prominence ratio	
Avg:	8.38 dB	Avg:	13.02 dB	Avg:	8.60 dB*	Avg:	16.34 dB*
Min:	7.56 dB*	Min:	9.59 dB	Min:	7.30 dB	Min:	10.34 dB*
Max:	9.46 dB	Max:	18.20 dB	Max:	10.01 dB*	Max:	24.25 dB*
SD:	0.73 dB	SD:	2.85 dB	SD:	1.03 dB*	SD:	4.09 dB*

Notes: See references [51, 52].
*denotes the greatest value per comparison.