Prediction of Alzheimer's disease progression within 6 years using speech: a novel approach leveraging language models

Samad Amini, MSc[a], Boran Hao, MSc[a], Jingmei Yang, MSc[a], Cody Karjadi, MSc[c], Vijaya B. Kolachalama, PhD[b,d,e], Rhoda Au, PhD[f,c], and Ioannis Ch. Paschalidis, PhD[a,d,g]

[a]Department of Electrical & Computer Engineering, Division of Systems Engineering, and Department of Biomedical Engineering, Boston University, 8 St. Mary's St, Boston, MA 02215

[b]Department of Medicine, Boston University School of Medicine, 72 E Concord St, Boston, MA 02118

[c]Framingham Heart Study, Boston University, 73 Mt Wayte Ave, Framingham, MA 01702

[d]Faculty of Computing & Data Sciences, Boston University, 665 Commonwealth Ave, Boston, MA 02215

[e]Department of Computer Science, Boston University, 665 Commonwealth Ave, Boston, MA 02215

[f]Departments of Anatomy & Neurobiology, Neurology, and Epidemiology, Boston University School of Medicine and School of Public Health, 72 E Concord St, Boston, MA 02118

[g]Corresponding author: Ioannis Ch. Paschalidis, yannisp@bu.edu, +1(617)694-8498, 8 St. Mary's St, Boston, MA 02215

Abstract

INTRODUCTION: Identification of individuals with Mild Cognitive Impairment (MCI) who are at risk of developing Alzheimer's Disease (AD) is crucial for early intervention and selection of clinical trials.

METHODS: We applied natural language processing techniques along with machine learning methods to develop a method for automated prediction of progression to AD within 6 years using speech. The study design was evaluated on the neuropsychological test interviews of n = 166 participants from the Framingham Heart Study, comprising 90 progressive MCI and 76 stable MCI cases.

RESULTS: Our best models, which used features generated from speech data, as well as age, gender, and education level, achieved an accuracy of 78.5% and a sensitivity of 81.1% to predict MCI-to-AD progression within 6 years.

DISCUSSION: The proposed method offers a fully automated procedure, providing an opportunity to develop an inexpensive, broadly accessible, and easy-to-administer screening tool for MCI-to-AD progression prediction, facilitating development of remote assessment.

Background

Alzheimer's disease (AD) is the most common cause of dementia and has a long prodromal phase, during which subtle cognitive changes occur. Mild Cognitive Impairment (MCI) is a stage between normal cognition and AD. Individuals with MCI are at higher risk of developing AD with a 3–15% conversion rate of MCI to AD every year [1,2]. Therefore, accurately predicting the progression of MCI to AD can assist physicians in making decisions regarding patient treatment, participation in cognitive rehabilitation programs, and selection for clinical trials involving new drugs [3].

Traditionally, AD pathology can be assessed using biomarkers such as cerebrospinal fluid assays or neuroimaging techniques like Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI) [4–7]. Several studies have explored these modalities to predict conversion from MCI to dementia [8–12]. Although these techniques provide useful information, they are invasive and expensive, limiting their applicability to well-resourced places and lacking the scalability and accessibility needed for low- and middle-income countries [13]. Furthermore, clinical and pathological variability is observed in AD using imaging techniques, which can make accurate diagnosis and prognosis challenging [14].

In contrast, a Neuro-Psychological Test (NPT), conducted through an in-person interview, is currently the most accessible method for assessing cognitive decline. The NPT, triggered by patient history and in conjunction with a clinical examination, provides a comprehensive evaluation of cognitive function, including attention, memory, language, and visuospatial abilities. Researchers have explored computer-based approaches to predict the progression from MCI to AD using NPTs [15–18], primarily relying on hand-crafted features and the cognitive scores extracted from the NPT by clinicians. However, these approaches have not yet achieved full automation, limiting their potential for more precise and efficient cognitive evaluations.

On the other hand, speech in the NPTs can be a strong predictor of cognitive decline [19,20], and various AI-assisted diagnostic models using linguistic and acoustic features extracted from the NPTs have been developed [21–23]. The Framingham Heart Study (FHS), which is the longest ongoing longitudinal, transgenerational cohort study of chronic disease, has been digitally recording the NPT interviews since 2005, and these voice recordings include all major established cognitive tests, such as the Boston Naming Test, Hooper Visual Organization Test, and Wechsler Memory Scale [24]. Several studies have used these recordings to develop diagnostic tools. For instance, a voice-based predictor was developed to identify dementia using acoustic features [25]. Xue et. al applied deep

learning methods to acoustic features from FHS voice recordings to detect dementia and MCI [26]. In our earlier work, we utilized Natural Language Processing (NLP) on the voice recordings to place each individual across the dementia spectrum [27].

NLP, particularly Large Language Models (LLMs) popularized with the introduction of ChatGPT, has emerged as a powerful tool in healthcare, showing reliable performance in various tasks [28–30]. By leveraging LLMs, we open up new frontiers in AD research, leading to the development of automated screening tools. Specifically, we consider the classification problem of determining whether individuals with MCI will progress to AD dementia within a 6-year window. Predicting conversions over a shorter period of time may be relatively easier, but has limited clinical utility [31].

Our automated pipeline utilizes audio recordings of the NPT to predict the likelihood of MCI subjects transitioning to AD within 6 years. We emphasize that our analysis only uses text automatically transcribed from these recordings and it does not rely on any acoustic features. By leveraging transformer-based language models, we aim to capture semantic nuances potentially missed by conventional scoring, enriching the assessment with comprehensive text features. This underscores our plan for developing a cost-effective, automated tool that surpasses traditional methods in detecting AD progression. Conducting the NPT interview remotely, via a web interface without clinician participation, can further minimize screening costs. The pipeline incorporates diverse computational techniques, including speech recognition, speech diarization, a transformer-based sentence encoder, and logistic regression models.

Methods

Study participants

A cohort of 166 subjects with cognitive complaints were consecutively monitored by the FHS [32], consisting of 59 males and 107 females, with a median age of 81 years (range: 63 to 97 years). It is noteworthy that the demographic composition of our cohort is predominantly White, reflecting the specific population from which the participants were drawn. Each participant underwent an approximately one-hour-long NPT, which was recorded and saved in the .wav format. The NPTs conducted by the FHS include sub-tests assessing different cognitive domains, such as memory, naming and language, visuoperceptual skills, abstract reasoning, and attention [33,34]. Additional information such as education, presence of Apolipoprotein E (ApoE) genes, and health risk factors (such

as blood glucose, diabetes, hypertension etc.) were also available. All the participants have a completed NPT for which an MCI diagnosis was assigned. The cognitive status assignments such as AD diagnosis and MCI for those showing signs of cognitive decline was reached by consensus of at least one neurologist and one neuropsychologist, based on neurology exams, FHS study and external medical records, and brain imaging (the diagnostic procedure is outlined [33,35]). All participants have provided written informed consent and study protocols and consent forms were approved by the Boston University Medical Campus Institutional Review Board.

Data preparation

The cohort for this study was derived from a larger group of participants whose NPTs were recorded by the FHS. This group consists of individuals at various cognitive stages, including some who have been diagnosed with MCI. Due to the increasing interest in AD and related clinical trials, our analysis focused on predicting the progression from MCI to AD. We elected not to consider progression from normal cognition to AD (or MCI) since the NPT has limited utility in predicting future cognitive decline in individuals without any current signs of cognitive deterioration. Therefore, we focused on MCI cases and identified those who had either progressed to AD or remained MCI within 6 years, as determined by a dementia review.

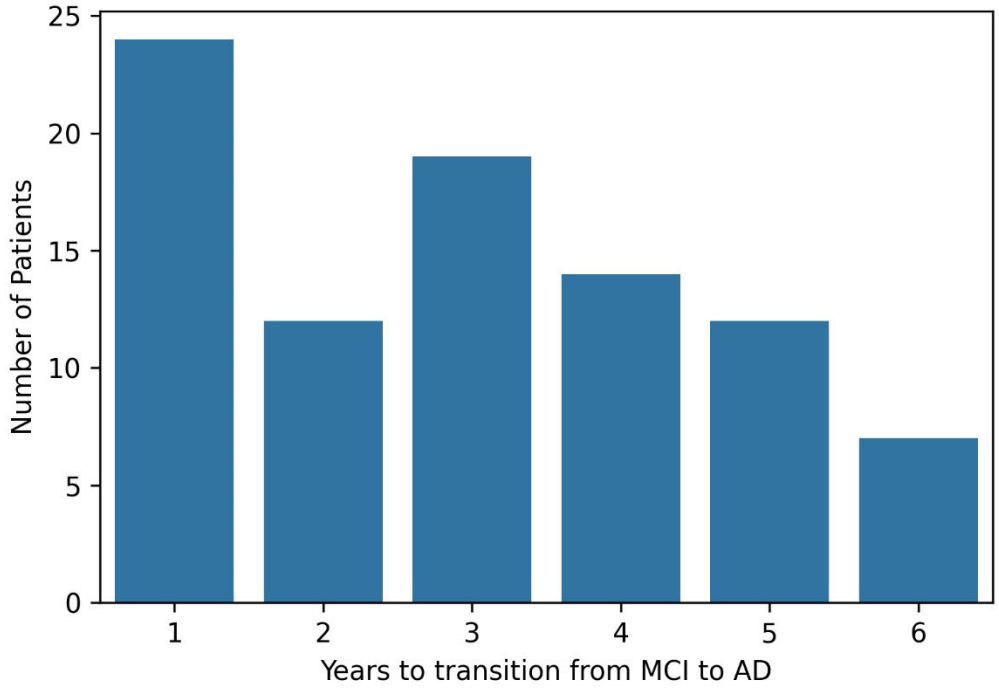


Figure *1* shows the number of patients transitioning to AD from MCI each year over this period. It represents the distribution of transitions, indicating that a larger number of patients tend to transition to AD earlier within the 6-year timeframe. This observation suggests that the progression from MCI to AD is more likely to occur in the initial years following the MCI diagnosis.

In our previous work [27], we developed a tool to automatically transcribe voice recordings. Each utterance was diarized (i.e., ascribed to a speaker: participant or examiner) and each transcript was split into the eight sub-tests comprising the FHS NPT. Some of these sub-tests are part of larger batteries of cognitive assessments such as Wechsler Memory Scale (WMS) [36], Wechsler Adult Intelligence Scale (WAIS) [37], and a revised form of the WAIS (WAIS-R) [38]. In addition, there are several other tests that are frequently administered independently, including the Boston Naming Test (BNT) [39], Verbal fluency (FAS) [40], and Clock Drawing Test (CDT) [41]. The other two sub-tests are DEMO, which represents a part of the interview related to demographic information, and OTHER, which includes parts that are not categorized in the defined sub-tests. Using this developed tool, the participants' audio files were automatically transcribed, and each sentence was automatically labeled based on the specific sub-test to which it belonged, such as WMS, WAIS, WAIS-R, BNT, FAS, CDT, DEMO, or OTHER.

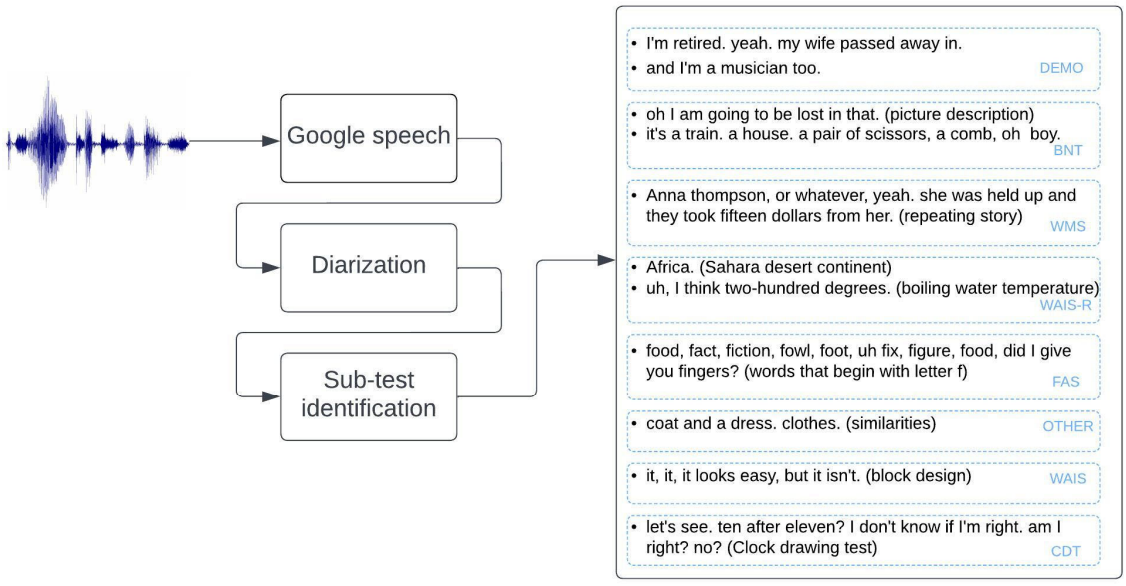


Figure *2* illustrates the automated pipeline to extract such structured data from the raw voice recording. From the prior study [27] leveraging a similar population, the diarization task demonstrated a performance with an Exact F1-score of 70.2%, and the sub-test classification task achieved an accuracy of 96.2%.

Statistical analysis

The cohort consisted of 166 patients with MCI, 90 of whom progressed to AD dementia (progressive MCI) and 76 remained MCI (stable MCI) within the 6-year horizon. AD dementia included AD with stroke, AD without stroke, and mixed dementia (vascular + AD). Over a 6-year follow-up period, the participants with MCI had a mean (s.d.) time to AD of 2.7 (1.5) years. Table 1 presents the participant characteristics, including self-reported gender, education status, age statistics, and six possible combinations of the three types of the ApoE gene (E2/E3/E4) for both copies of the allele. The table suggests that older women with lower education levels and those carrying one or two copies of the ApoE E4 allele are more likely to progress to AD. This finding aligns with previous studies that highlight age as the most significant risk factor for AD [42]. As individuals age, the prevalence of AD increases significantly, with estimates of 19% for those aged 75-84 and 30-35% for those over 85 years old [43]. Additionally, research shows that individuals who inherit one copy of the ApoE E4 genotype have a higher risk of developing AD, while those who inherit two copies have an even higher risk [44,45]. Notably, in the progressive MCI group, females had an average age of 1.4 years older than males, suggesting that females may be more prone to progression due to their longer lifespan.

Transcript encoding using Universal Sentence Encoder

There are currently no standard methods for encoding a document into quantitative data. Based on selecting a specific segment of each transcript, we obtain different vector embeddings for each NPT interview. To increase the training data, we randomly sample from each transcript to create several abbreviated versions that are then encoded. In addition, the content of each sub-test can be encoded separately, resulting in 8 specific embeddings. These embedding vectors are generated by a deep learning-based model, the Universal Sentence Encoder (USE) [46]. The USE is a pre-trained neural network based on the transformer architecture and has demonstrated a promising downstream classification accuracy on dementia detection and other tasks [27,47]. The USE outputs a 512-dimensional vector for each embedding. To simplify the downstream classification model, we perform dimensionality reduction using a logistic regression-based Recursive Feature Elimination (RFE) method [48]. Specifically, we perform logistic regression-based RFE on the training data, systematically removing the weakest feature as determined by the smallest absolute value of the logistic regression coefficients.

Prediction procedure

We generate deep learning-based embedding vectors from either an abbreviated version of a transcript or the content of one specific sub-test. This results in 8 embedding vectors associated with each sub-test, as well as multiple embedding vectors from the abbreviated versions of one transcript. We then train a logistic regression model on the quantitative data associated with one sub-test content, resulting in 8 different trained models and 8 scores for the sub-tests. However, the 8 scores representing the sub-tests undergo a feature selection process using performance error analysis. The embeddings from multiple shortened versions of each transcript are treated as independent input, and one logistic regression model is trained on all of them, resulting in the generation of multiple scores for one transcript. Although the abbreviated versions of a transcript are treated independently during the embedding procedure, we take the average of the logistic regression scores to create the Transcript Average Score (TAS). Finally, we feed the TAS score along with the selected sub-test scores into an ensemble logistic regression model to make the final prediction of the likelihood of an individual with MCI converting to AD within 6 years.

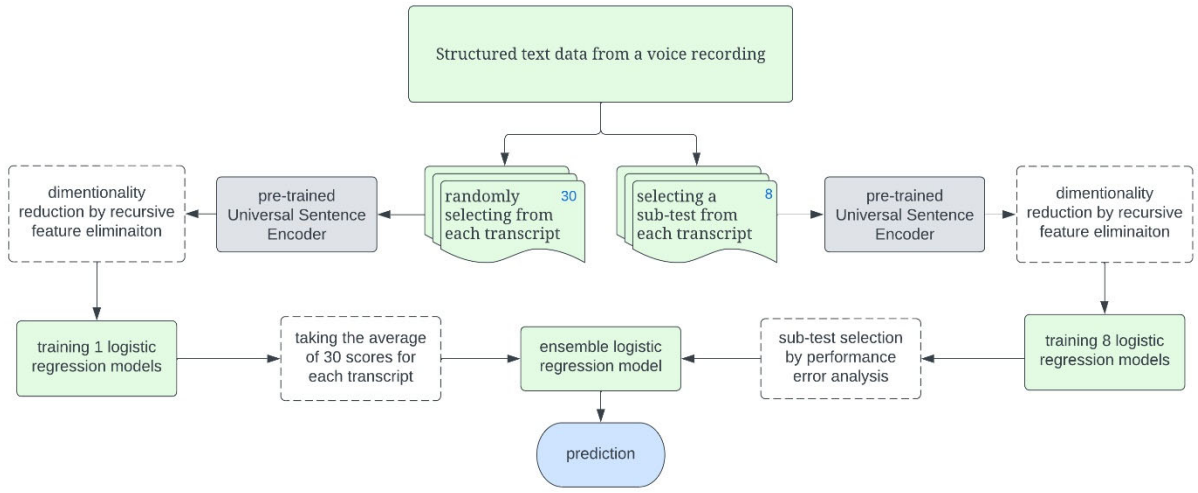
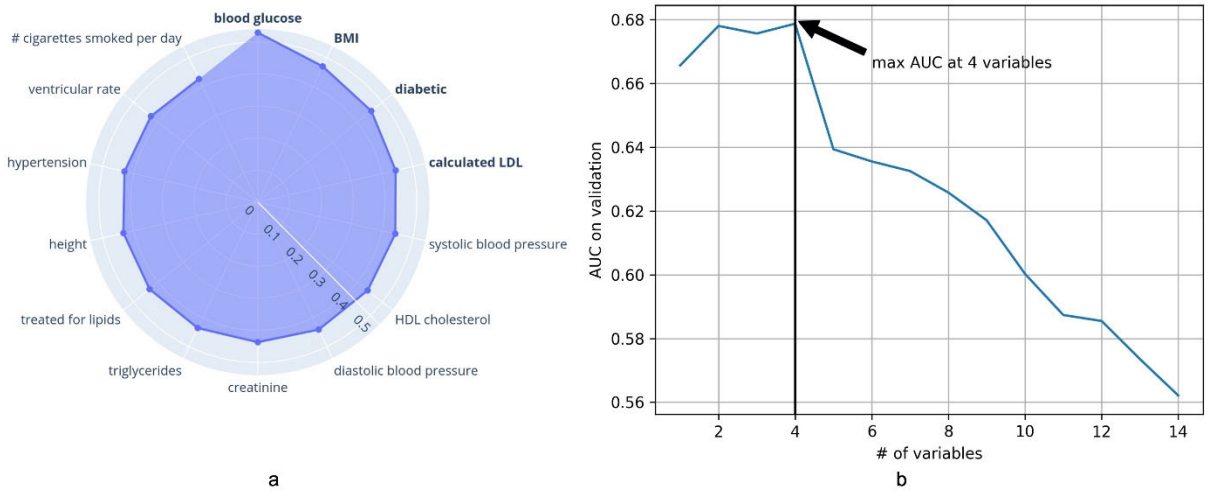


Figure *3*: Automated pipeline for Alzheimer's disease prediction from an NPT interview.



illustrates the prediction process. By integrating random abbreviation and sub-test specific embeddings through data augmentation, our approach significantly enhances the model's data interpretation and accuracy. This includes generating the TAS score from diverse transcript versions, alongside sub-test evaluations to improve our prediction process. This strategy enriches our model's data representation and predictive accuracy, leveraging both broad and detailed transcript insights.

Validation and performance metrics

To evaluate our model's performance, we employed a stratified group k-fold cross-validation approach, splitting the dataset into 10 folds. This division allocated 90% of the data for training (across 9 folds) and 10% for testing (the remaining fold), with each segment serving as the test set once to ensure comprehensive evaluation. Within this framework, we also implemented an internal cross-validation within the training phase for dimensionality reduction and feature selection. This nested cross-validation strategy ensures the test data remain unseen until the final testing phase, enhancing the validity and reliability of our results. We conducted the stratified group k-fold cross-validation three times, each with a distinct random seed, to accurately calculate the average metrics and 95% confidence intervals for our model's performance assessment. The performance metrics considered for the evaluation were classification accuracy, sensitivity, specificity, precision, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC). The AUC is a valuable measure that estimates the probability of the classifier ranking a randomly chosen progressive MCI subject (positive sample) higher than a randomly selected

stable MCI subject (negative sample). Sensitivity and specificity provide insights into the correct classification of positive and negative subjects, while the F1 score measures the trade-off between precision and recall.

Results

Table 2 presents the average performance metrics of the logistic regression model, including the 95% confidence interval for each metric. The table is sorted in descending order based on AUC, with the highest value listed first. The first row showcases the model's performance, incorporating text, demographics, ApoE, and health factors, achieving an AUC of 78.5% and an F1 score of 79.9%, marking the highest effectiveness observed. The subsequent two rows highlight models that leverage text features along with readily available demographic data such as age, sex, and education, also demonstrating strong predictive capabilities with an AUC and F1-socre of 77.8% and 79.4% for our NLP model using only text features. The fourth row of the table reports the performance of adding ApoE data to the model using demographic features, resulting in an AUC and F1-score of 71.7% and 75.7%. In addition, we trained a model with only demographic features as input, yielding an AUC of 68.8% as shown in row 6th.

We also assessed a logistic regression model based on traditional neuropsychological test scores, including assessments like Logical Memory, Visual Reproductions, Paired Associate Learning Immediate Recall, Similarity Test, Boston Naming Test, and Verbal Fluency Test. The model's performance, detailed in the 5th row, shows an AUC of 71.3% and an F1-score of 75.5%, underscoring that our NLP model not only matches but exceeds the predictive power of standard NPT scores. Additionally, when using 4 health factors (blood glucose, body mass index, presence of diabetes, and calculated low-density lipoprotein (LDL)) as input to the logistic regression, the seventh row shows an AUC of 66.2% and F1 score of 72.5%. As MMSE evaluates cognitive problems with thinking, communication, understanding, and memory, the model based on MMSE yielded an AUC of 60.7%. Other combinations of different features had no performance improvement over the best models in the first three rows of Table 2. Furthermore, the 4 health factors used in Table 2 (blood glucose, body mass index, presence of diabetes, and calculated LDL) resulted from the performance error analysis of 14 health factors; see the Supplement and Figure 4 for the complete analysis.

Based on the confidence intervals detailed in Table 2, the performance metrics of the first three rows, which utilize the text feature set, distinguish them significantly from other models presented in the table. While there

may be some overlap in confidence intervals between models using text features and baseline models, statistical analysis, such as the paired t-test, validates that the AUC for models employing text features is significantly improved, underscoring the efficacy of our NLP approach in enhancing predictive accuracy.

Figure 5[1] displays the coefficients of our logistic regression model using the text features and the demographics model output. The results have been adjusted for continuous variables through z-score normalization (by subtracting the mean and dividing by the standard deviation), making the coefficients comparable. This figure represents the distribution of logistic regression coefficients for different features, highlighting their relative importance in the model's predictive process. By comparing the interquartile ranges and medians of coefficients for TAS and selected sub-tests against the demographic features, we can observe a difference in their contributions. A higher median value for TAS and sub-tests implies these variables have a stronger predictive value, underscoring their role over demographic factors in influencing the model's prediction.

Discussion

Speech during cognitive exams has been identified as a promising biomarker that strongly correlates with underlying cognitive dysfunction. The current study aimed to automatically predict the progression to AD using NLP and machine learning techniques applied to speech data. The proposed method predicted the participant's progression to AD with an accuracy of 78.2% and a sensitivity of 81.1% in the held-out test data, demonstrating strong predictive power over a 6-year span. However, the specificity of predicting whether an individual with MCI will progress to AD within 6 years was moderate, at 75%. To reduce the costs associated with recruiting subjects for clinical trials, it is important to improve the specificity. Nevertheless, the relatively high sensitivity of our prediction tool makes it clinically applicable and potentially beneficial for eventual neuroprotective therapies [49].

Importantly, our method only utilizes features derived from speech data in an automated manner, along with easily obtainable variables such as age, gender, and education level. The proposed method offers a non-invasive, accessible, and easy-to-administer AI-based predictive approach because it does not require data involving laboratory tests, genetic tests, or imaging exams. This makes it a promising candidate for integration into remote

[1] demographics: age, sex, and education; TAS: transcript average score; BNT: Boston Naming Test; DEMO: part of the interview related to demographic information; WAIS: Wechsler Adult Intelligence Scale; CDT: Clock Drawing Test; OTHER: similarity tests.

assessment technologies. A major strength of this study is its utilization of semantic features extracted from the structured text data. This approach allows for the potential transferability of the entire pipeline to other languages, leveraging the availability of transcription tools that can transcribe from any language to English, and/or powerful NLP models in different languages [50,51]. As a computer-aided decision-making tool, our method has the potential to mitigate inter-clinician variability in selecting candidates for clinical trials and drug tests, enhancing the consistency and reliability of participant selection processes [52].

The Results section indicates that adding demographic features to text features does not enhance the model's ability to predict the progression from MCI to AD. This contrasts with previous assumptions about the predictive power of age and other demographics in relation to degenerative diseases over extended periods. Even though there are significant differences in demographics between stable and progressive MCI groups, the use of text features alone outperforms the use of demographic features. This underscores the strong predictive strength of the engineered text features. Moreover, upon evaluating the performance of the logistic regression model using the traditional NPT scores, we observed an AUC of 71.3%. This result indicates that our approach outperforms conventional NPT scoring methods in this study. Furthermore, when we compared our model with a well-established cognitive assessment score such as the MMSE score, text features demonstrated higher predictive power. In addition, compared with other works that used only non-invasive features [53,54], our model's F1-score = 79.4% is higher. For instance, the authors in one paper [53] predicted AD transition within 9 years based on NPT scores provided by specialized clinicians, achieving an F1-score of 70.8%, whereas M. Grassi et al. achieved an F1-score of 72.7% using socio-demographic characteristics, clinical information, and NPT scores [54]. These methods still require highly specialized personnel to generate the NPT scores while our method is fully automated, making Alzheimer's prediction accessible to all.

As depicted in Figure 5, our analysis revealed that sub-tests related to demographic questions (DEMO), Boston Naming Test (BNT), similarity tests (OTHER), and Wechsler Adult Intelligence Scale (WAIS) emerged as the top features driving the performance of our model. These sections of each transcript are key predictors for identifying the future incidence of AD. Thus, our approach facilitates the identification of sub-tests that provide more informative insights for predicting the future incident of AD. This finding underscores the potential benefit of employing a more structured interview to better capture the language deficits that may underlie cognitive decline. Additionally, after conducting a performance error analysis on 14 health risk factors, we found that variables such as

blood glucose, BMI, diabetes, and calculated LDL were useful in predicting the development of Alzheimer's disease. In conclusion, our study demonstrates the potential of using automatic speech recognition and NLP techniques to develop a prediction tool for identifying individuals with MCI who are at risk of developing AD. Our method achieved high accuracy and outperformed other non-invasive approaches. However, further prospective studies with larger populations are necessary to validate the generalizability of our models. Additionally, it is important to standardize the definition of MCI across different locations to enable better comparison of results. With continued development and refinement, our approach may contribute to early intervention and selection in clinical trials for novel AD treatments, ultimately improving patient outcomes.

References

[1] Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. Acta Psychiatr Scand 2009;119:252–65.

[2] Liu S, Cao Y, Liu J, Ding X, Coyle D, Initiative ADN. A novelty detection approach to effectively predict conversion from mild cognitive impairment to Alzheimer's disease. Int J Mach Learn Cybern 2023;14:213–28.

[3] Pereira T, Ferreira FL, Cardoso S, Silva D, De Mendonça A, Guerreiro M, et al. Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease: a feature selection ensemble combining stability and predictability. BMC Med Inform Decis Mak 2018;18:1–20.

[4] Scheltens P, Blennow K, Breteler M, de Strooper B, Frisoni G, Salloway S, et al. Alzheimer's disease. Lancet (Lond Engl) 388: 505–517 2016.

[5] Turner RS, Stubbs T, Davies DA, Albensi BC. Potential new approaches for diagnosis of alzheimer's disease and related dementias. Front Neurol 2020;11:496.

[6] Thomas JA, Burkhardt HA, Chaudhry S, Ngo AD, Sharma S, Zhang L, et al. Assessing the Utility of Language and Voice Biomarkers to Predict Cognitive Impairment in the Framingham Heart Study Cognitive Aging Cohort Data. J Alzheimers Dis 2020;76:905–22.

[7] Weiner MW, Veitch DP, Miller MJ, Aisen PS, Albala B, Beckett LA, et al. Increasing participant diversity in AD research: Plans for digital screening, blood testing, and a community-engaged approach in the Alzheimer's Disease Neuroimaging Initiative 4. Alzheimers Dement 2023;19:307–17.

[8] Caminiti SP, Ballarini T, Sala A, Cerami C, Presotto L, Santangelo R, et al. FDG-PET and CSF biomarker accuracy in prediction of conversion to different dementias in a large multicentre MCI cohort. NeuroImage Clin 2018;18:167–77.

[9] Long X, Chen L, Jiang C, Zhang L, Initiative ADN. Prediction and classification of Alzheimer disease based on quantification of MRI deformation. PloS One 2017;12:e0173372.

[10] Varatharajah Y, Ramanan VK, Iyer R, Vemuri P. Predicting short-term MCI-to-AD progression using imaging, CSF, genetic factors, cognitive resilience, and demographics. Sci Rep 2019;9:2235.

[11] Ahmadzadeh M, Christie GJ, Cosco TD, Moreno S. Neuroimaging and analytical methods for studying the pathways from mild cognitive impairment to Alzheimer's disease: protocol for a rapid systematic review. Syst Rev 2020;9:1–6.

[12] Ritter K, Schumacher J, Weygandt M, Buchert R, Allefeld C, Haynes J-D, et al. Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. Alzheimers Dement Diagn Assess Dis Monit 2015;1:206–15.

[13] Clute-Reinig N, Jayadev S, Rhoads K, Le Ny A-L. Alzheimer's disease diagnostics must be globally accessible. J Alzheimers Dis 2021;84:1453–5.

[14] Kelley S, Perez-Urrutia N, Morales R. Misfolded amyloid-β strains and their potential roles in the clinical and pathological variability of Alzheimer's disease. Neural Regen Res 2023;18:119.

[15] Tabert MH, Manly JJ, Liu X, Pelton GH, Rosenblum S, Jacobs M, et al. Neuropsychological prediction of conversion to Alzheimer disease in patients with mild cognitive impairment. Arch Gen Psychiatry 2006;63:916–24.

[16] Chapman RM, Mapstone M, McCrary JW, Gardner MN, Porsteinsson A, Sandoval TC, et al. Predicting conversion from mild cognitive impairment to Alzheimer's disease using neuropsychological tests and multivariate methods. J Clin Exp Neuropsychol 2011;33:187–99.

[17] Silva D, Guerreiro M, Santana I, Rodrigues A, Cardoso S, Maroco J, et al. Prediction of long-term (5 years) conversion to dementia using neuropsychological tests in a memory clinic setting. J Alzheimers Dis 2013;34:681–9.

[18] Pereira T, Lemos L, Cardoso S, Silva D, Rodrigues A, Santana I, et al. Predicting progression of mild cognitive impairment to dementia using neuropsychological data: a supervised learning approach using time windows. BMC Med Inform Decis Mak 2017;17:1–15.

[19] Stück D, Signorini A, Alhanai T, Sandoval M, Lemke C, Glass J, et al. Novel Digital Voice Biomarkers of Dementia from the Framingham Study. Alzheimers Dement 2018;14:P778–9.

[20] Boschi V, Catricala E, Consonni M, Chesi C, Moro A, Cappa SF. Connected speech in neurodegenerative language disorders: a review. Front Psychol 2017;8:269.

[21] Hernández-Domínguez L, Ratté S, Sierra-Martínez G, Roche-Bergua A. Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. Alzheimers Dement Diagn Assess Dis Monit 2018;10:260–8.
[22] Liu L, Zhao S, Chen H, Wang A. A new machine learning method for identifying Alzheimer's disease. Simul Model Pract Theory 2020;99:102023.
[23] Pulido MLB, Hernández JBA, Ballester MÁF, González CMT, Mekyska J, Smékal Z. Alzheimer's disease and automatic speech analysis: A review. Expert Syst Appl 2020;150:113213.
[24] Downer B, Fardo DW, Schmitt FA. A summary score for the Framingham Heart Study neuropsychological battery. J Aging Health 2015;27:1199–222.
[25] Lin H, Karjadi C, Ang TF, Prajakta J, McManus C, Alhanai TW, et al. Identification of digital voice biomarkers for cognitive health. Explor Med 2020;1:406.
[26] Xue C, Karjadi C, Paschalidis IC, Au R, Kolachalama VB. Detection of dementia on raw voice recordings using deep learning: A Framingham Heart Study. Available SSRN 3788945 2021.
[27] Amini S, Hao B, Zhang L, Song M, Gupta A, Karjadi C, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach. Alzheimers Dement 2022.
[28] Iroju OG, Olaleke JO. A systematic review of natural language processing in healthcare. Int J Inf Technol Comput Sci 2015;7:44–50.
[29] Srivastava SK, Singh SK, Suri JS. A healthcare text classification system and its performance evaluation: A source of better intelligence by characterizing healthcare text. Cogn. Inform. Comput. Model. Cogn. Sci., Elsevier; 2020, p. 319–69.
[30] Robin J, Xu M, Balagopalan A, Novikova J, Kahn L, Oday A, et al. Characterizing progressive speech changes in prodromal-to-mild Alzheimer's disease using natural language processing. Alzheimers Dement 2022;18.
[31] Chen J, Chen G, Shu H, Chen G, Ward BD, Wang Z, et al. Predicting progression from mild cognitive impairment to Alzheimer's disease on an individual subject basis by applying the CARE index across different independent cohorts. Aging 2019;11:2185.
[32] Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS. 70-year legacy of the Framingham Heart Study. Nat Rev Cardiol 2019;16:687–98.
[33] Au R, Piers RJ, Devine S. How technology is reshaping cognitive assessment: Lessons from the Framingham Heart Study. Neuropsychology 2017;31:846.
[34] Jak AJ, Preis SR, Beiser AS, Seshadri S, Wolf PA, Bondi MW, et al. Neuropsychological criteria for mild cognitive impairment and dementia risk in the Framingham Heart Study. J Int Neuropsychol Soc JINS 2016;22:937.
[35] Satizabal CL, Beiser AS, Chouraki V, Chêne G, Dufouil C, Seshadri S. Incidence of dementia over three decades in the Framingham Heart Study. N Engl J Med 2016;374:523–32.
[36] Wechsler D, Scale-Revised WAI. The psychological corporation. San Antonio TX 1997.
[37] Wechsler D. The measurement and appraisal of adult intelligence, 1958. Baltim Wiliams Wilkins 2020.
[38] Zarantonello MM, Munley PH, Milanovich J. Predicting wechsler adult intelligence scale-revised (WAIS-R) IQ scores from the luria-nebraska neuropsychological battery (form I). J Clin Psychol 1993;49:225–33.
[39] Goodglass H, Kaplan E. The assessment of aphasia and related disorders. Lea & Febiger; 1972.
[40] Franzen MD. Multilingual aphasia examination. Kans City MO Test Corp Am 1986.
[41] Amini S, Zhang L, Hao B, Gupta A, Song M, Karjadi C, et al. An artificial intelligence-assisted method for dementia detection using images from the clock drawing test. J Alzheimers Dis 2021;83:581–9.
[42] Herrup K. Reimagining Alzheimer's disease—an age-based hypothesis. J Neurosci 2010;30:16755–62.
[43] Armstrong RA. Risk factors for Alzheimer's disease. Folia Neuropathol 2019;57:87–105.
[44] Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis. Jama 1997;278:1349–56.
[45] Liu C-C, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat Rev Neurol 2013;9:106–18.
[46] Cer D, Yang Y, Kong S, Hua N, Limtiaco N, John RS, et al. Universal sentence encoder. ArXiv Prepr ArXiv180311175 2018.
[47] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. ArXiv Prepr ArXiv170603762 2017.
[48] Hao B, Sotudian S, Wang T, Xu T, Hu Y, Gaitanidis A, et al. Early prediction of level-of-care requirements in patients with COVID-19. Elife 2020;9:e60519.

[49] Eskildsen SF, Coupé P, García-Lorenzo D, Fonov V, Pruessner JC, Collins DL, et al. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. Neuroimage 2013;65:511–21.
[50] Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, et al. Multilingual universal sentence encoder for semantic retrieval. ArXiv Prepr ArXiv190704307 2019.
[51] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised cross-lingual representation learning at scale. ArXiv Prepr ArXiv191102116 2019.
[52] Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, Meilán JJG. Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: a systematic review article. Front Psychol 2021;12:620251.
[53] Merone M, D'Addario SL, Mirino P, Bertino F, Guariglia C, Ventura R, et al. A multi-expert ensemble system for predicting Alzheimer transition using clinical features. Brain Inform 2022;9:20.
[54] Grassi M, Rouleaux N, Caldirola D, Loewenstein D, Schruers K, Perna G, et al. A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures. Front Neurol 2019;10:756.

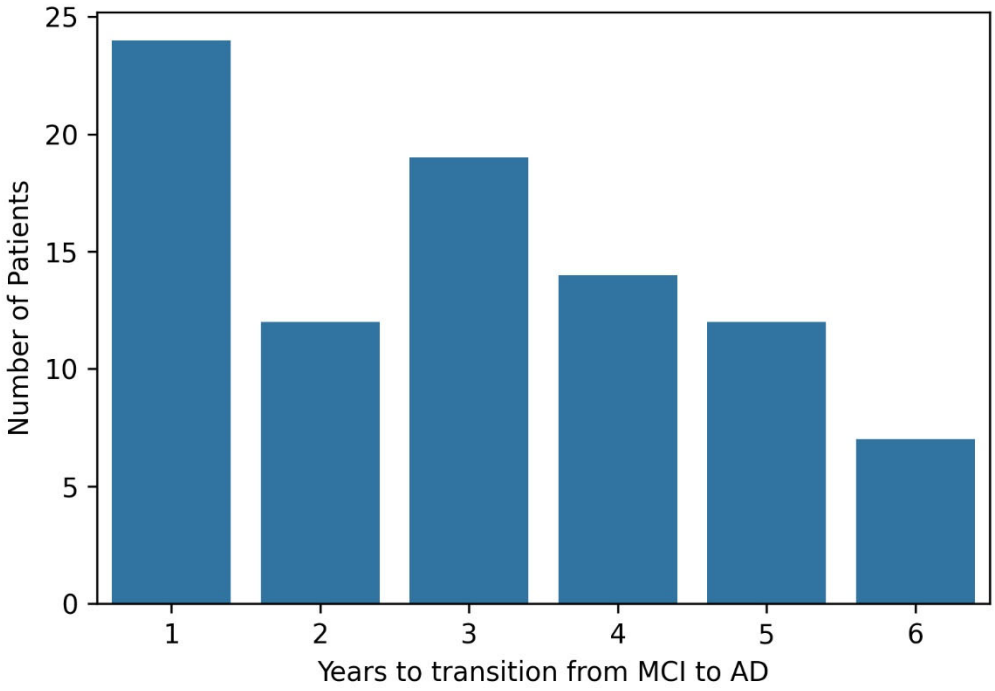


Figure 1: Number of MCI patients transitioning to AD annually over 6 years.

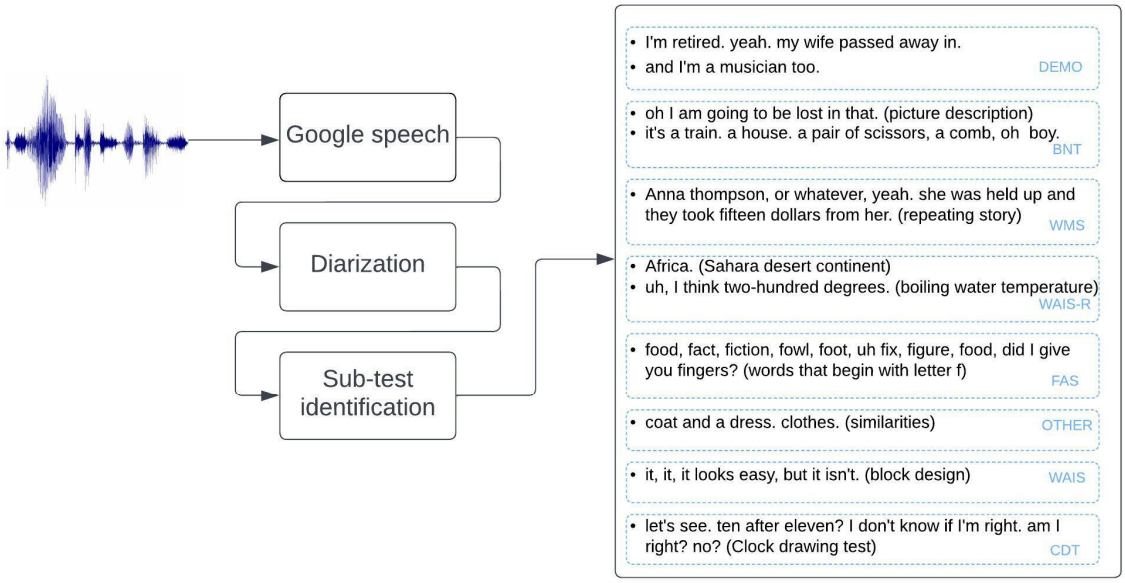


Figure 2: Automated pipeline for converting raw speech into structured data (as an example, the box on the right side contains a short note from each sub-test highlighted in blue ink).

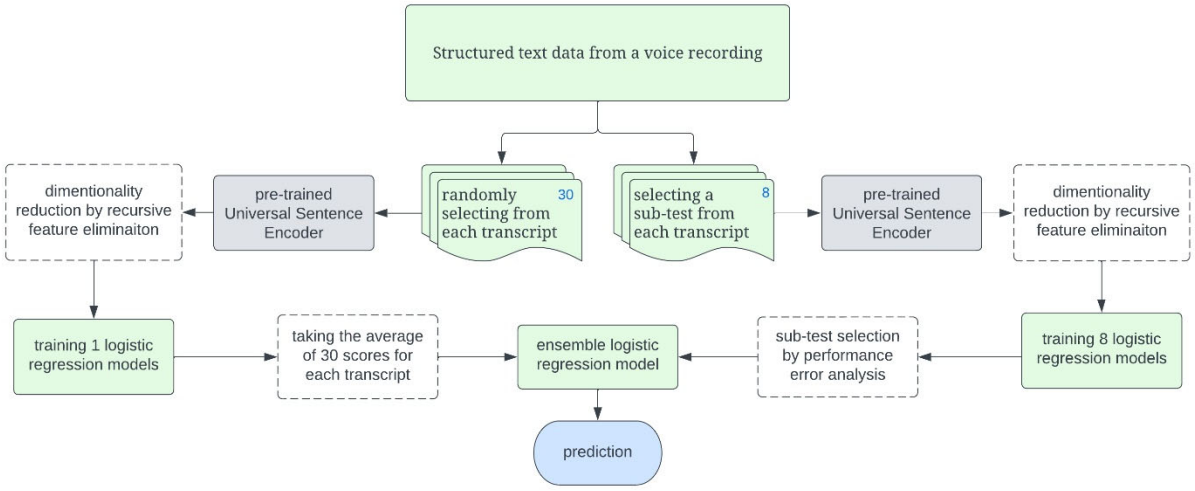


Figure 3: Automated pipeline for Alzheimer's disease prediction from an NPT interview.

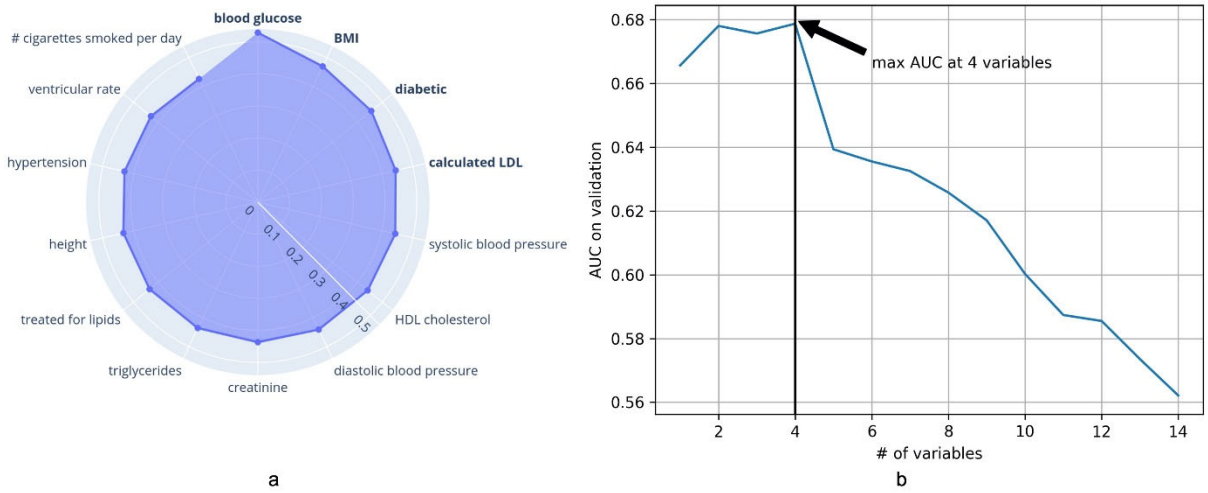


Figure 4: Performance error analysis for health factors. (a) performance error (1-AUC) after removing each feature at a time. (b) results of AUC for an arbitrary number of most important features.

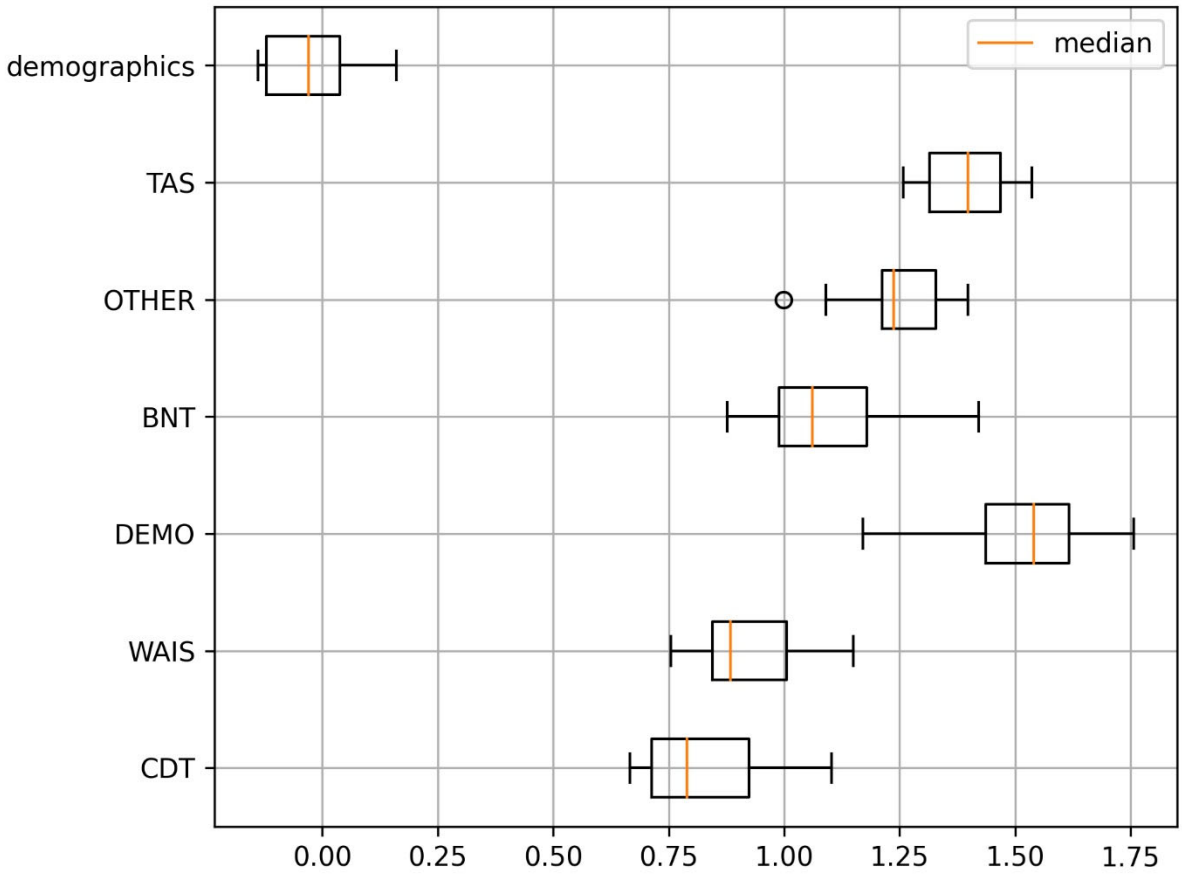


Figure 5: Logistic regression coefficients of the text features and demographics used in the proposed method.

Funding

The research was partially supported by the NSF under grants CCF-2200052, DMS-1664644, and IIS-1914792, the NIH under grants R01 GM135930 and UL54 TR00413, and the Boston University Rajen Kilachand Fund for Integrated Life Science and Engineering.

Conflicts

Rhoda Au is a scientific advisor to Signant Health and NovoNordisk and consultant to Biogen and the Davos Alzheimer's Collaborative. She receives funding from the National Institute on Aging (AG072654, AG062109, AG068753) and has also been supported through awards from the American Heart Association, the Alzheimer's Drug Discovery Foundation, Alzheimer's Disease Data Initiative and Gates Ventures. Vijaya B. Kolachalama has received support from the Karen Toffler Charitable Trust, Johnson & Johnson (through the Boston

University Lung Cancer Alliance), the NIH under grants RF1-AG062109, R01-HL159620, R43-DK134273, and R21-CA253498, the American Heart Association under grant 20SFRN35460031, and serves as a consultant to AstraZeneca. Both R. Au and V. B. Kolachalama state no conflicts of interest with the present work. There is no declaration from other authors.

Consent Statement

All participants have provided written informed consent and study protocols and consent forms were approved by the Boston University Medical Campus Institutional Review Board.

Key Words

Alzheimer's disease prognosis, Natural Language Processing, Neuropsychological test, Framingham Heart Study

Table 1: Characteristics of patients with MCI, who either remain MCI or progress to AD within 6 years.

	stable MCI $n = 76$	**progressive MCI** $n = 90$	Difference
Age			
63-75	29	8	-21
75-85	36	44	not significant
85+	11	38	27
Gender [mean age]			
Female	44 [77.8]	63 [84.2]	19
Male	32 [77]	27 [82.8]	not significant
Education			
High school grad or less	33	46	13
Some college or more	43	44	not significant
ApoE			
E44	1	6	5
E34 or E24	19	29	10
E22, E33, or E23	52	54	not significant

Table 2: Average performance metrics (over 30 runs) on a held-out test set of the final logistic regression models using different features for MCI-to-AD progression in 6 years. Acc.: Accuracy, Sens.: Sensitivity, Spec.: Specificity, Prec.: Precision.

Features	**AUC**	**Acc.**	**Sens.**	**Prec.**	**Spec.**	**F1-score**
text & demographics & Apoe & health	78.5 (74.6, 82.5)	78.8 (75.6, 82.1)	80.6 (75.9, 85.2)	80.4 (76.8, 84.1)	76.9 (72.2, 81.5)	79.9 (74.6, 82.5)

text	77.8 (74.2, 81.3)	78.2 (75.0, 81.4)	81.1 (75.9, 86.3)	78.9 (75.8, 81.9)	75.0 (70.9, 79.1)	79.4 (76.0, 82.7)
text & demographics	77.5 (73.8, 81.2)	78.5 (75.4, 81.7)	81.1 (75.8, 86.3)	79.3 (76.1, 82.6)	75.6 (71.4, 79.8)	79.6 (76.1, 83.0)
demographics & Apoe	71.7 (67.7, 75.6)	74.4 (71.9, 76.9)	77.8 (71.5, 84.1)	77.1 (73.3, 80.9)	70.6 (63.6, 77.6)	75.7 (72.7, 78.7)
traditional NP tests	71.3 (67.2, 75.5)	74.7 (71.8, 77.6)	77.2 (70.7, 83.7)	77.2 (73.5, 80.8)	71.9 (66.3, 77.5)	75.5 (72.0, 79.0)
demographics	68.8 (64.3, 73.3)	70.6 (67.1, 74.1)	70.6 (64.5, 76.6)	74.9 (70.4, 79.4)	70.6 (64.3, 77.0)	71.1 (67.5, 74.8)
health factors	66.2 (63.1 71.2)	71.2 (68.2, 74.1)	75.0 (68.4, 81.6)	73.2 (69.7, 76.7)	66.9 (61.2, 72.5)	72.5 (68.9, 76.1)
MMSE	60.7 (55.9, 65.4)	62.9 (59.5, 64.4)	66.7 (60.8, 72.6)	65.2 (61.4, 69.0)	58.8 (52.9, 64.6)	64.9 (61.1, 68.8)

Prediction of Alzheimer's disease progression within 6 years using speech: a novel approach leveraging language models

Supplementary

Samad Amini, MSc[a], Boran Hao, MSc[a], Jingmei Yang, MSc[a], Cody Karjadi, MSc[c], Vijaya B. Kolachalama, PhD[b,d,e], Rhoda Au, PhD[f,c], and Ioannis Ch. Paschalidis, PhD[a,d,g]

[a]Department of Electrical & Computer Engineering, Division of Systems Engineering, and Department of Biomedical Engineering, Boston University, 8 St. Mary's St, Boston, MA 02215

[b]Department of Medicine, Boston University School of Medicine, 72 E Concord St, Boston, MA 02118

[c]Framingham Heart Study, Boston University, 73 Mt Wayte Ave, Framingham, MA 01702

[d]Faculty of Computing & Data Sciences, Boston University, 665 Commonwealth Ave, Boston, MA 02215

[e]Department of Computer Science, Boston University, 665 Commonwealth Ave, Boston, MA 02215

[f]Departments of Anatomy & Neurobiology, Neurology, and Epidemiology, Boston University School of Medicine and School of Public Health, 72 E Concord St, Boston, MA 02118

[g]Corresponding author: Ioannis Ch. Paschalidis, yannisp@bu.edu, +1(617)694-8498, 8 St. Mary's St, Boston, MA 02215

We extended our model training and validation to include two additional time windows representing the progression of MCI patients to AD. In the 5-year time frame, we had a cohort of 111 stable and 83 progressive samples, while in the 7-year time frame, we had 47 stable and 98 progressive samples. The results of these analyses can be found in Table S1. Notably, we observed that the model utilizing generated text features continued to outperform the model relying on demographics and ApoE genotype information.

To select the 4 health factors used in the models reported in Table 2, we first conducted correlation analysis using the Pearson method among 23 health factors. A total of 9 variables were removed due to a strong positive correlation with the rest. The remaining 14 features included blood glucose, body mass index (BMI), calculated low-density lipoprotein (LDL) cholesterol, number of cigarettes smoked per day, creatinine, history of diabetes, average diastolic blood pressure, height, history of hypertension, whether or not treated for lipids, average systolic blood pressure, triglycerides, and ventricular rate per minute. We then performed additional feature selection on the remaining 14 features using performance error analysis. This involved training a logistic regression model with all 14 features, and then systematically removing one feature at a time from the training set and evaluating the performance of the model each time. The features were then ranked based on their relative performance error. As shown in Figure 4: Performance error analysis for health factors. (a) performance error (1-AUC) after removing each feature at a time. (b) results of AUC for an arbitrary number of most important features.

4(a), blood glucose emerged as the most important feature with the highest error among other health factors, and the rest of the parameters were sorted based on their decreasing performance error in a clockwise order on the radar chart. Given these findings, we determined that using only the 4 most important variables from the health factors yielded the best performance outcome, according to Figure 4: Performance error analysis for health factors. (a) performance error (1-AUC) after removing each feature at a time. (b) results of AUC for an arbitrary number of most important features.

4(b).

Table S1: Performance metrics (average over 10 runs) on a held-out test set of the final logistic regression models for MCI-to-AD progression in 5 and 7 years.

Features	Features Time window, years	AUC	Accuracy	F1 score
Text & demographics	5	73.6	75	71
	7	76.8	76.7	79.8
Text	5	72.7	75.5	70.8
	7	75.2	74.7	77.7
demographics & Apoe	5	63.5	67.5	67.4
	7	68.5	71.3	75