Detection of dementia on raw voice recordings using deep learning: A Framingham Heart Study

Chonghua Xue[1,*], Cody Karjadi[1,2,*], Ioannis Ch. Paschalidis[3], Rhoda Au[2,4], Vijaya B. Kolachalama[1,5,§]

[1]Section of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, MA, USA - 02118
[2]The Framingham Heart Study, Boston University, Boston, MA, USA - 02118
[3]Departments to Electrical & Computer Engineering, Systems Engineering and Biomedical Engineering; Faculty of Computing & Data Sciences, Boston University, Boston, MA, USA - 02118
[4]Departments of Anatomy & Neurobiology and Neurology, Boston University School of Medicine, Boston, MA, USA - 02118; Boston University Alzheimer's Disease Center, Boston, MA, USA - 02118; Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA - 02118
[5]Department of Computer Science and Faculty of Computing & Data Sciences, Boston University, Boston, MA, USA - 02115; Boston University Alzheimer's Disease Center, Boston, MA, USA - 02118

***Indicates equal contribution**

[§]Corresponding author:

Vijaya B. Kolachalama, PhD

72 E. Concord Street, Evans 636, Boston, MA - 02118

Email: vkola@bu.edu

Phone: (+1) 617-358-7253

ORCID: https://orcid.org/0000-0002-5312-8644

Twitter: @vkola_lab

Abstract

Background: Identification of reliable, affordable and easy-to-use strategies for detection of dementia are sorely needed. Digital technologies, such as individual voice recordings, offer an attractive modality to assess cognition but methods that could automatically analyze such data without any pre-processing are not readily available.

Methods: We used a subset of 1264 digital voice recordings of neuropsychological examinations administered to participants from the Framingham Heart Study (FHS), a community-based longitudinal observational study. The recordings were 73 minutes in duration, on average, and contained at least two speakers (participant and clinician). Of the total voice recordings, 483 were of participants with normal cognition (NC), 451 recordings were of participants with mild cognitive impairment (MCI), and 330 were of participants with dementia. We developed two deep learning models (a two-level long short-term memory (LSTM) network and a convolutional neural network (CNN)), which used the raw audio recordings to classify if the recording included a participant with only NC or only dementia, and also to differentiate between recordings corresponding to non-demented (NC+MCI) and demented participants.

Findings: Based on 5-fold cross-validation, the LSTM model achieved a mean (±std) area under the sensitivity-specificity curve (AUC) of 0.744±0.038, mean accuracy of 0.680±0.032, mean sensitivity of 0.719±0.112, and mean specificity of 0.652±0.089 in predicting cases with dementia from those with normal cognition. The CNN model achieved a mean AUC of 0.805±0.027, mean accuracy of 0.740±0.033, mean sensitivity of 0.735±0.094, and mean specificity of 0.750±0.083 in predicting cases with only dementia from those with only NC. For the task related to classification of demented participants from non-demented ones, the LSTM model achieved a mean AUC of 0.659±0.043, mean accuracy of 0.701±0.057, mean sensitivity of 0.245±0.161 and mean specificity of 0.856±0.105. The CNN model achieved a mean AUC of 0.730±0.039, mean accuracy of 0.735±0.046, mean sensitivity of 0.443±0.113, and mean specificity of 0.840±0.076 in predicting cases with dementia from those who were not demented.

Interpretation: This proof-of-concept study demonstrates the potential that raw audio recordings of neuropsychological testing performed on individuals recruited within a community cohort setting can provide a level of screening for dementia.

Introduction

Impairment in cognition is a common manifestation among all the individuals suffering from dementia, of which Alzheimer's disease (AD) is the most common. Despite the rising dementia pandemic accompanying a rapidly aging population, there remains a paucity of cognitive assessment tools that are applicable regardless of age, education, language/culture. Thus, there is an urgent need to identify reliable, affordable and easy-to-use strategies for detection of signs of dementia. If these approaches can also detect disease earlier than currently available tests, then this can significantly impact clinical management of dementia.

Starting in 2005, the Framingham Heart Study (FHS) cognitive aging studies began digitally recording all spoken responses to neuropsychological tests, allowing precise capture of all spoken error and extraneous responses for tests such as paragraph recall, naming and verbal fluency; this accurate documentation allowed for novel application of the Boston Process Approach (BPA),[1] a scoring method to potentially detect subtle changes in NP function that cannot be captured using typical scoring practices. Comprehensive and accurate transcription of the entirety of the recordings requires manual transcription because automated speech recognition tools are still inaccurate. Further, this manual work is not only tedious, but also requires specialized expertise in neuropsychology, speech recognition, natural language understanding and speech translation. Such expertise is not readily available. Developing a computational tool that can simply take a raw voice recording as an input and automatically assess the dementia status of the individual has broad implications for dementia screening tools that are scalable across diverse populations.

Promising solutions to tackle such datasets may be found in the field of deep learning, an approach to data analysis that has increasingly been used over the past few years to address an array of formerly intractable questions in medicine. Deep learning,[2] a subfield of machine learning, is based upon specific models known as neural networks which decompose the complexities of observed datasets into hierarchical interactions among low-level input features. Once these interactions are learned, refined and formalized by exposure to a training dataset, fully-trained deep learning models leverage their "experience" of prior data to make predictions about new cases. Thus, these approaches offer powerful medical decision-making potential due to their ability to rapidly identify low-level signatures of disease from large datasets, and quickly apply them at scale. This hierarchical approach makes deep learning ideally suited to derive novel insights from high volumes of audio/voice data. Our primary objective was to develop a long short-term memory network (LSTM) model and a convolutional neural network (CNN) model, to predict dementia status on the FHS raw voice recordings, without any manual processing of the audio content. As a secondary objective, we processed the model-derived salient features and reported the distribution of time spent by individuals on various neuropsychological tests and these tests' relative contribution to diagnosis.

Methods

Study population

The voice data consists of digitally recorded neuropsychological examinations administered to participants from the Framingham Heart Study (FHS), which is a community-based longitudinal observational study that was initiated in 1948 and consists of several generations of participants.[3] The neuropsychological examinations consist of multiple tests that assess memory, attention, executive function, language, reasoning, visuoperceptual skills, and premorbid intelligence. Further information and details can be found in Au et. al.,[4] including the lists of all the neuropsychological tests included in each iteration of the FHS neuropsychological battery.[5]

Cognitive status determination

The cognitive status of the participants over time was diagnosed via the FHS dementia diagnostic review panel.[6,7] The panel consists of at least one neuropsychologist and at least one neurologist. The panel reviews neuropsychological and neurologica exams, medical records, and family interviews for each participant. Selection for dementia review is based on whether participants have shown evidence of cognitive decline, as has been previously described.[4]

The panel creates a cognitive timeline for each participant that provides the participant's cognitive status on a given date over time. In order to label the participants' cognitive statuses at the time of each recording, we selected the closest date of diagnosis to the recording that fell on or before the date of recording or after the recording within 180 days. If the closest date of diagnosis was more than 180 days after the recording, but the participant was determined to be cognitively normal, we labeled that participant as cognitively normal. Dementia diagnosis was based on criteria from the *Diagnostic and Statistical Manual of Mental Disorders,* fourth edition (DSM-IV) and the NINCDS-ADRDA for Alzheimer's dementia.[8] The study was approved by the Institutional Review Boards of Boston University Medical Center and all participants provided written consent.

Digital voice recordings

FHS began to digitally record the audio of neuropsychological examinations in 2005. Overall, FHS has 9786 digital voice recordings from 5449 participants. The dataset for this study includes digital voice recordings from September 2005 to March 2020 from the subset of participants who came to dementia review, either because they were tagged as potentially cognitively impaired or had died. Since participants only go to dementia review after showing signs of cognitive impairment or having passed away, some of the cognitively normal participants are biased because they were suspected to be cognitively impaired but review revealed they were not. This contrasts with the larger majority of participants deemed cognitively normal individuals who were not brought to review prior to death, and presumably remained cognitively normal during the entire life. Because repeat recordings were available on some participants, there are also some recordings from a participant in which at the initial recording they were deemed

cognitively intact, and then at the time of a subsequent recording, they had become impaired. Thus, they would be still represented as cognitively intact for the time frame prior to their date of progression to dementia diagnosis. This dataset includes 656 unique participants with 1264 recordings, which are 73.13 (±22.96) minutes in duration on average. The range is from 8.43 to 210.82 minutes. There are 483 recordings of participants with normal cognition (291 participants), 451 recordings of participants with mild cognitive impairment (309 participants), 934 recordings of non-demented participants (507 participants), and 330 recordings of demented participants (223 participants) **(Tables 1 & S1)**. Of the 656 unique participants, one participant may have several recordings with different cognitive statuses. For example, one participant could have normal cognition at their first recording, mild cognitive impairment at their second recording, and dementia at their third recording. As noted above, participants with a recording associated with either normal cognition, MCI, or dementia are not mutually exclusive and will not necessarily add up to the overall 656 unique participants. The recordings are all in the WAV format and were downsampled to 8 kHz.

Data preprocessing

To preserve as much useful information as possible, the Mel-frequency cepstral coefficients (MFCCs) were extracted during the data preprocessing stage. (MFCCs) are the coefficients that collectively make up the Mel-frequency cepstrum, which serves as an important acoustic feature in many speech processing applications. Leveraging the nonlinear Mel scaling makes the response much closer to the human auditory system and therefore renders it an ideal feature for speech-related tasks. Each FHS audio recording was first split into short frames with a window size of 25 ms (i.e., 200 sample points at 8000Hz sampling rate) and a stride length of 10 ms. For each frame, the periodogram estimate of the power spectrum was calculated by 256-point discrete Fourier transformation. Then, a filterbank of 26 triangle filters evenly distributed with respect to the Mel scale was applied to each frame. We then applied a discrete Cosine transformation of the logarithm of all filterbank energies. Note that the 26 filters correspond to 26 coefficients, but in practice, only 2nd-13th coefficients are believed to be useful; we loosely followed this convention while replacing the first coefficient with total energy that might contain helpful information on the entire frame.

Hierarchical long short-term memory network model

We developed a two-level hierarchical long short-term memory network (LSTM) model to associate the raw voice recordings with dementia status (**Figure 1A**). The LSTM model's unique architecture is particularly suitable to tackle long sequences of data.[9] A one-hour recording may yield hundreds of thousands of temporally ordered MFCCs vectors, while the memory capacity of the LSTM model is empirically limited to only a few thousand vectors. We first grouped every 2000 consecutive MFCC vectors into segments without overlap. The low-level LSTM was used to process these segments one by one and output a new vector for each of them. The collection of the new vectors formed another sequence, which was then fed into the high-level

LSTM to generate a single vector summarizing the entire recording. Note that the hierarchical design ensured that both the two-level LSTM architecture was not overwhelmed by longer sequences beyond their practical limitation. For the last step of the LSTM scheme, a multilayer perceptron (MLP) network was used to estimate the probability of dementia based on the summarized vector.

Convolutional neural network model

For comparison with the LSTM model, we designed a one-dimensional convolutional neural network (CNN) model for dementia classification (**Figure 1B**). The model consisted of 7 stacked convolutional blocks. Within each block, there were 2 convolutional layers, 1 max-pooling layer and 1 nonlinearity function. The convolutional layer was set to have kernel length 3, stride length 1 and zero padding. The max-pooling layer had both kernel length and stride length fixed at 4. The nonlinearity function was one from the ReLU family. Note that the max-pooling layer would reduce the number of input vectors by factor of 4; by stacking 7 of them, the number of the output vectors resulting from the very top convolutional block could be 256 times less than the number of input MFCCs vectors such that all the relative information of an entire audio signal was highly compressed. A global average pooling layer was then introduced to aggregate all the output vectors by averaging across the temporal axis. This step ensured that all information was contained within a fixed length vector regardless of the duration of any given recording and based on which logistic regression layer could estimate the probability of dementia.

Saliency maps

We derived the saliency maps based on the intermediate result right before the global average pooling step of the CNN model. The intermediate result was composed of two vectors which signified DE+ and DE- prediction respectively. For simplicity, we only used the DE+ vector. Since the recording-level prediction was determined by the average of the saliency map which also preserved temporal structure, it allowed us to observe finer aspects of the prediction by revealing the values assigned to each short period. For our CNN model settings, each value corresponded to roughly two minutes and a half of an original recording. Note that the length of the periods is implied by the CNN model parameters; altering the stride size, the kernel size, the number of stacked CNN blocks etc. may result in different lengths. To align the saliency map with the original recording for further analysis, we extended its size to the total number of seconds via nearest neighbor interpolation.

Salient administered fractions

The CNN model tasked with classifying between participants with normal cognition and dementia was tested on 81 participants who have transcripts for 123 recordings out of their 183 total recordings. Those 60 recordings without transcripts were excluded from the saliency analysis but were included in the test dataset. Of the recordings with transcripts, there were 44 recordings of participants with normal cognition and 79 recordings of participants with dementia.

The recordings that were transcribed were divided into subgroups based on the time spent by the participant on each neuropsychological test. As a result, for each second of a given recording, the DE[+] saliency value and the current neuropsychological test are known. In order to calculate the salient administered fraction (SAF) for a given neuropsychological test, we counted the number of seconds for that neuropsychological test that also had a DE[+] saliency value greater than zero and then divided it by the total number of seconds for that neuropsychological test. For example, if the Boston Naming Test (BNT) has 90 seconds that had DE[+] saliency values greater than zero and BNT was administered for 100 seconds, then the SAF[+] would be 0.90. Similarly, we produced SAF[-] in the same way, except for periods of time where the DE[+] saliency value was less than or equal to zero. The SAF[+] was calculated for every administered neuropsychological test for each recording that had a demented participant and that the model classified the participant as demented (true positive). The SAF[-] was calculated in the same way, except for recordings with participants with normal cognition and the model classified to have normal cognition (true negative).

Statistical analysis

The models were constructed on a workstation with a GeForce RTX 2080 Ti graphics processing unit. The models were trained using 5-fold cross-validation. Performance was presented as mean and standard deviation over the model runs. We generated sensitivity-specificity (SS) and precision-recall (PR) curves based on the cross-validated model predictions. For each SS and PR curve, we also computed the area under curve (AUC) values. Additionally, we computed sensitivity, specificity, F1-score and Matthews correlation coefficient on the cross-validated model predictions.

Role of the funding source

The study sponsors had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

We observed that the FHS participants on average spent different amounts of time to complete specific neuropsychological tests, and these times varied as a function of their cognitive status for most of the tests (**Figure 2**). For example, almost all participants spent the highest amount of time while completing the Boston Naming Test (BNT). During this test, demented (DE) participants spent significantly more time (611.06±260.25 s) than the non-demented (NDE) participants (390.67±167.41 s). This observation was also valid when the time spent by DE participants was compared specifically with those who had normal cognition (NC) (405.91±176.78 s), or with the ones who had mild cognitive impairment (MCI) (321±93 s). However, there was no statistical significance between the times taken for BNT by NC and MCI participants. A similar pattern was observed for a few other neuropsychological tests including 'Visual Reproductions Delayed Recall' (NC: 140.61±73.22 s, MCI: 114.78±28.31 s, DE: 79.13±48.17 s and NDE: 135.96±67.88 s), 'Verbal Paired Associates Recognition' (NC: 75.79±21.62 s, MCI: 90.00±26.42 s, DE: 126.04±52.24 s and NDE: 78.29±22.92 s), 'Copy Clock' (NC: 70.10±33.54 s, MCI: 54.67±13.97 s, DE: 90.91±44.40 s and NDE: 67.01±31.13 s), 'Trails A' (NC: 115.42±58.26 s, MCI: 105.33±23.71 s, DE: 198.97±107.08 s and NDE: 113.49±53.32 s), and 'Trails B' (NC: 219.24±107.57 s, MCI: 241.00±117.50 s, DE: 431.23±278.04 s and NDE: 223.40±108.55 s). We also observed no statistically significant differences in the times taken for the NC, MCI, DE and NDE participants on a few other neuropsychological tests including 'Logical Memory Immediate Recall' (NC: 134.98±28.56 s, MCI: 116.90±19.46 s, DE: 128.87±38.50 s and NDE: 131.57±27.84 s), 'Digit Span Forward' (NC: 115.91±31.40 s, MCI: 100.10±32.32 s, DE: 107.47±36.67 s and NDE: 112.98±31.87 s), 'Similarities' (NC: 227.78±89.64 s, MCI: 217.33±79.52 s, DE: 227.99±120.45 s and NDE: 225.94±87.28 s), 'Verbal Fluency (FAS)' (NC: 325.50±38.97 s, MCI: 330.00±29.23 s, DE: 336.10±65.36 s and NDE: 326.31±37.18 s), 'Finger Tapping' (NC: 294.40±106.89 s, MCI: 323.80±87.64 s, DE: 289.00±130.62 s and NDE: 300.28±102.32 s), 'Information (WAIS-R)' (NC: 392.65±181.18 s, MCI: 335.75±156.89 s, DE: 372.72±153.56 s and NDE: 385.07±176.72 s), 'Cookie Theft' (NC: 198.93±75.28 s, MCI: 244.63±101.69 s, DE: 186.94±146.98 s and NDE: 206.54±79.58 s) and 'Miscellaneous' (NC: 140.15±199.84 s, MCI: 183.50±199.77 s, DE: 111.66±119.14 s and NDE: 151.46±196.20 s).

Both the LSTM and the CNN models that were trained and validated on the FHS voice recordings demonstrated consistent performance across the different data splits used for 5-fold cross validation (**Figure 3**). For the classification of demented versus normal participants, the LSTM model took 138 minutes to fully train (8 epochs and batch size of 4) and took 14 seconds to predict on a test case; the CNN model took 490 minutes to fully train (32 epochs and batch size of 4) and took 13 seconds to predict on a test case. For the classification of demented versus non-demented participants, the LSTM model took 437 minutes to fully train (16 epochs and batch size of 4) and took 20 seconds to predict on a test case; the CNN model took 778 minutes

to fully train (32 epochs and batch size of 4) and took 20 seconds to predict on a test case. In general, we observed that the CNN model performed better on most metrics for both sets of classification problems, i.e., NC versus DE (**Table 2A**), as well as DE versus NDE (**Table 2B**). The only exceptions were when the LSTM model sensitivity was higher for the case of NC versus dementia classification, and the LSTM model specificity was higher for the demented versus non-demented case classification.

We computed the average (± standard deviation) of SAF[+] and SAF[-] derived from the CNN model (**Table 3**). The positive SAFs were calculated for all true positive recordings and the negative SAFs were calculated for all true negative recordings. For example, the SAF[+] for the 'Verbal paired associates recognition' test was 0.88±0.26. This indicates that on average 88% of the duration of the 'Verbal paired associates recognition' tests administered in true positive recordings also intersected with segments of time that the model found to be DE[+] salient. Also, the SAF[-] for the 'Verbal paired associates recognition' test was 0.32±0.42. This indicates that on average 32% of the duration of the 'Verbal paired associates recognition' tests administered in true negative recordings also intersected with segments of time that the model found to not be DE[+] salient. On the other hand, the SAF[+] for the 'Command clock' test was 0.39±0.45, indicating that only about 39% of the duration of the "Command clock' tests administered in true positive recordings also intersected with segments of time that the model found to be DE[+] salient. Also, the SAF[-] for the 'Command clock' test was 0.76±0.37. The rest of the average positive SAFs and average negative SAFs were also reported for the remaining neuropsychological tests as well as the number of true positive or true negative recordings that contained an administration of the given test (**Table 3**). We also computed average SAF[+] and SAF[-] for additional neuropsychological tests (**Table S1)** which were set aside due to the low number of samples. A schematic of the DE[+] saliency vectors used to generate SAF[+] and SAF[-] can be seen in **Figure 4**. Each value in a given saliency vector represented approximately 2 minutes and 30 seconds of a recording. Since the saliency vector covers the entire recording, each second of every neuropsychological test within a recording can be assigned a saliency vector value, which was then used to calculate SAF[+] and SAF[-].

Discussion

Cognitive performance is affected by numerous factors that are independent of underlying actual cognitive status. Physical impairments (vision, hearing), mood (depression, anxiety), low education, cultural bias, as well as test-specific characteristics are just a few of many variables that can lead to variable performance. Further, neuropsychological tests that claim to test the same specific cognitive domains (e.g., memory, executive function, language, etc.) do not do so in a uniform manner. For example, when testing verbal memory, a paragraph recall test taps into a different set of underlying cognitive processes compared to that of a list learning task. Also, the presumption that a test measures only a single cognitive domain is naively simplistic since every neuropsychological test involves multiple cognitive domains to complete. Thus, across many factors, there is significant heterogeneity of cognitive performance that make it difficult to differentiate between those who will become demented versus those who will not, especially at the preclinical stage. Particularly, traditional paper and pencil neuropsychological tests may not be sufficiently sensitive to pick up when subtle change begins. While cognitive complaints serve as a surrogate preclinical measure of decline, there is the inherent bias of self-report. Given this complex landscape, digital voice recordings of neuropsychological tests provide a data source of relative independence from the aforementioned limitations. To our knowledge, our study is the first to demonstrate that a continuous stream of data is also amenable for automated analysis for evaluation of cognitive status.

Digital health technologies in general and voice in particular, are increasingly being evaluated as potential screening tools for depression,[10-13] and various neurodegenerative diseases such as Parkinson's disease.[14-17] Recently, potential opportunities for developing digital biomarkers based on mobile/wearables for Alzheimer's disease were outlined.[18,19] Our study is unique in focusing on two deep learning methods that relies on a completely hands-free approach for processing raw voice recordings to predict dementia status. The advantage of our approach is three-fold. The first is the limited need to extensively process any of the voice recordings before sending them as inputs to the neural networks. This is a major advantage because it minimizes the burden of generating accurate transcriptions and/or handcrafted features that generally take time to develop and relying on the availability of experts who are not necessarily readily available. This aspect places us in a unique position compared to previously published work that depended on derived measures.[20,21] Second, our approach can process audio recordings of variable lengths, which means that one does not have to format or select a specific window of the audio recording for analysis. This important strength underscores the generalizability of our work because one can process voice recordings containing various combinations of neuropsychological tests that are not bounded within a time frame. Finally, our approach allows for identification of audio segments that are highly correlated with the outcome of interest. The advantage of doing this is that it provides a "window" into the machine learning black box; we can go back to the recordings and identify the various speech patterns or segments of the

neuropsychological tests, which point to high probability of disease risk and understand their contextual significance.

The CNN architecture allowed us to generate the saliency vectors by utilizing the parameters of the final classification layer. Simply put, a temporal saliency vector for each specific case could be obtained by calculating the weighted sum of the output feature vectors from the last convolutional block in the CNN, indicating how each portion of the recording contributed to either positive or negative prediction. We then aligned the saliency vectors with the recording timeline to further analyze the speech signatures to understand if there were any snippets of the neuropsychological testing that often were correlated with the output class label. From examining the transcriptions that exist for a portion of the dataset, we were able to identify which neuropsychological tests were occurring during any given time in a recording, and then calculated the positive SAFs.

All the FHS voice recordings contained audio content from two distinct speakers of which one is the interviewee (participant), and the other is the interviewer (clinician). We did not attempt to discern speaker-specific audio content as our models processed the entire raw audio recording at once. This choice was intentional because we wanted to first evaluate if deep learning can predict dementia status of the participant without having to perform detailed feature engineering on the audio recordings. Future work could focus on processing these signals and recognizing speaker-related differences and interactions, periods of silence and other nuances to make the audio recordings more amenable for deep learning. Also, additional studies can be performed to integrate the audio data with other routinely collected information that requires no additional processing (i.e., demographics) to augment model performance. An important point to note is that studies as proposed above need to be conducted with the goal of creating scalable solutions across the globe, particularly to those regions where technical or advanced clinical expertise is not readily available. This means that users at the point-of-care may not be in the best position to manually process raw voice recordings or any other data to derive needed features that can be fed into the computer models. Since our deep learning models do not require data pre-processing or handcrafted features, our approach has the ability to serve as a potential screening tool without having the need of an expert-level input. Our current findings serve as a first step towards achieving such a solution that can have a broader outreach.

Our study has a few limitations. The models were developed using the FHS voice recordings, which is a single population cohort from the New England area in the United States. Despite demonstrating consistent model performance using rigorous cross-validation (5-fold) approaches, our models still need to be validated using data from external cohorts to confirm their generalizability. Currently, we do not have access to any other cohort that has voice recordings of neuropsychological exams. Therefore, our study findings need to be interpreted in light of this limitation and with the hope of evaluating them further in the future. Due to lack of

available data in some cases and because not all participants took all the types of neuropsychological tests, we were able to generate the distribution of times spent for a portion of the tests. It must be also noted that the number of NC, MCI, DE and NDE participants varied for each neuropsychological test. Additionally, there may be outside factors affecting the amount of time it took a participant to complete a neuropsychological test that are not represented in the boxplots. For example, a normal participant could finish the BNT quickly and perform well, whereas administration of the BNT to a demented participant could be abruptly stopped due to an inability to complete the test or for other reasons. Therefore, the amount of time spent administering the BNT in the recordings in those two cases could be similar, but for different reasons. Nevertheless, statistical tests were performed to quantify the pairwise differences on all the available neuropsychological exams, which gave us the flexibility to report the differences that were statistically different and those that were similar.

In conclusion, our proposed deep learning approaches (LSTM & CNN) to processing raw voice recordings in an automated fashion allowed us to classify dementia status. Such approaches that rely minimally on neuropsychological expertise, audio transcription or manual feature engineering can pave the way towards the development of real-time screening tools in dementia care, especially in resource-limited settings.

Data sharing: All results, code and documentation will be made available on GitHub (https://github.com/vkola-lab/dvoice) upon manuscript publication. Data in this study cannot be shared publicly due to regulations of local ethical committees. Data might be made available to researchers who meet the criteria (to be provided once all data are available) for access to confidential data and upon Institutional Review Board's approval.

Funding: This project was supported in part by the Karen Toffler Charitable Trust, National Center for Advancing Translational Sciences, National Institutes of Health (NIH), through BU-CTSI Grant (1UL1TR001430), a Scientist Development Grant (17SDG33670323) and a Strategically Focused Research Network (SFRN) Center Grant (20SFRN35460031) from the American Heart Association, and a Hariri Research Award from the Hariri Institute for Computing and Computational Science & Engineering at Boston University, Framingham Heart Study's National Heart, Lung and Blood Institute contract (N01-HC-25195; HHSN268201500001I) and NIH grants (R01-AG062109, R21-CA253498, R01-AG008122, R01-AG016495, R01-AG033040, R01-AG054156, R01-AG049810, U19 AG068753). Additional support was provided by Boston University's Affinity Research Collaboratives program, Boston University Alzheimer's Disease Center (P30-AG013846), the National Science Foundation under grants DMS-1664644, CNS-1645681, and IIS-1914792, the Office of Naval Research under grant N00014-19-1-2571, and the NIH under grant R01-GM135930.

Conflict of interest: None

Author contributions: CX contributed to conceptualization, formal analysis, investigation, methods, validation, and visualization. CK contributed to data curation and verification, formal analysis, investigation, methods, validation and visualization. IP contributed to conceptualization and validation. RA contributed to the conceptualization, data curation and verification, validation and supervision. VBK contributed to overall supervision, conceptualization, investigation, validation, visualization, and writing the original draft. All authors contributed to writing the article and editing and have approved the final manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Main figure and table captions

Table 1: Demographics and participant characteristics. For each participant, digital voice recordings of neuropsychological examinations were collected. Tables (A), (B) and (C) show the demographics of the participants with normal cognition, mild cognitive impairment and dementia, respectively at the time of the voice recordings. Here, N represents the number of unique participants. The mean age (± standard deviation) is reported at the time of the recordings. Mean MMSE scores (± standard deviation) were computed closest to the time of the voice recording. For cognitively normal participants, ApoE data was unavailable for one Generation 1 (Gen 1) participant, and eight Generation (Gen) 2 participants; MMSE data was not collected on Generation (Gen) 3 participants. For MCI participants, ApoE data was unavailable for one Gen 1 participant, six Gen 2 participants, and one New Offspring Cohort (NOS) participant; MMSE data was also not collected for OmniGen2, and NOS participants and was not available for one Gen 1 participant. For demented participants, ApoE data was unavailable for six Gen 1 participants and three Gen 2 participants; MMSE data was not collected for Gen 3, OmniGen2, and NOS Cohort participants and not available for one Gen 1 participant.

Figure 1: Schematics of the deep learning frameworks. (A) Hierarchical long short-term memory (LSTM) network model that encodes an entire audio file into a single vector to predict dementia status on the individuals. All LSTM cells within the same row share the parameters. Note that the hidden layer dimension is user-defined (e.g., 64 in our approach.) (B) Convolutional neural network that uses the entire audio file as the input to predict the dementia status of the individual. Each convolutional block reduces the input length by a common factor (e.g., 2) while the very top layer aggregates all remaining vectors into a single one by averaging them.

Figure 2: Time spent on the neuropsychological tests. Boxplots showing the time spent by the FHS participants on each neuropsychological test. For each test, the boxplots were generated on participants with normal cognition (NC), those with mild cognitive impairment (MCI), and those who had dementia (DE); the no dementia group (NDE) combined the NC and MCI individuals. We also indicated the number of recordings that were processed to generate each boxplot. We also computed pairwise statistical significance between two groups (NC vs MCI, MCI vs DE, NC vs DE and DE vs NDE). The symbol '*' indicates statistical significance at $p<0.05$, the symbol '**' indicates statistical significance at $p<0.01$, the symbol '***' indicates statistical significance at $p<0.001$, and 'n.s.' indicates $p>0.05$. Logical Memory (LM) tests with a (†) symbol denote that a prompt was used to help facilitate recall when the participant was unable to do so unprompted. It is possible that one participant may receive a prompt under each of the LM recall conditions (one recording). Because a large number of neuropsychological tests were administered on the participants, we chose a representation scheme that combined colors and hatches. The colored hatches were used to represent each individual neuropsychological test and

this information was used to aid visualization in subsequent figures. This colored hatch key is referenced in subsequent figures.

Figure 3: Sensitivity-specificity (SS) and precision-recall (PR) curves of the deep learning models. The LSTM and CNN models were constructed to classify normal and demented participants as well as non-demented and demented participants, respectively. On each model, a 5-fold cross-validation was performed to generate the SS and PR curves. Plots (A) and (B) denote the SS and PR curves for the LSTM and the CNN models for the classification of normal versus demented cases. Plots (C) and (D) denote the SS and PR curves for the LSTM and CNN models for the classification of non-demented versus demented cases.

Table 2: Performance of the deep learning models. The LSTM and CNN models were constructed to classify (A) normal and demented participants as well as (B) non-demented and demented participants, respectively. The tables show the 5-fold cross-validation performance (mean ± standard deviation) of the LSTM and CNN models, respectively.

Figure 4: Saliency maps highlighted by the CNN model. (A) This key is a representation that maps the colored hatches to the neuropsychological tests. (B) Saliency map representing a recording (62 minutes in duration) of a participant with normal cognition (NC) that was classified as NC by the CNN model. (C) Saliency map representing a recording (94 minutes in duration) of a participant with dementia who was classified as demented by the CNN model. For both (B) & (C), the colormap on the left half corresponds to a neuropsychological test. The color on the right half represents the DE[+] value, ranging from dark blue (low DE[+]) to dark red (high DE[+]). Each DE[+] rectangle represents roughly 2 minutes and 30 seconds.

Table 3: Salient administered fractions derived from the CNN model. The average SAF (salient administered fraction) and standard deviation for true positive (SAF[+]) and true negative (SAF[-]) cases are listed in descending order based on the SAF[+] value. SAF[+] is calculated by summing up the time spent in a given neuropsychological test that intersects with a segment of time that is DE[+] salient and dividing by the total time spent in a given neuropsychological test. SAF[-] is calculated by summing up the time spent in a given neuropsychological test that intersects with a segment of time that is not DE[+] salient and dividing by the total time spent in a given neuropsychological test. The number of samples for SAF[+] and SAF[-] indicate the number of true positive and true negative recordings that contain each neuropsychological test.

Supplementary table captions

Table S1: Demographics of the participants who were non-demented (i.e., individuals with normal cognition and mild cognitive impairment) at the time of the voice recordings. ApoE data

was unavailable for one Gen 1 participant, thirteen Gen 2 participants, and one New Offspring Cohort (NOS) participant; MMSE data was not collected for all Gen 3, OmniGen 2, and NOS participants.

Table S2: For the neuropsychological tests that have too few samples, the average SAF (salient administered fraction) and standard deviation for true positive (SAF[+]) and true negative (SAF[-]) cases are listed in descending order based on the SAF[+] value. SAF[+] is calculated by summing up the time spent in a given neuropsychological test that intersects with a segment of time that is DE[+] salient and dividing by the total time spent in a given neuropsychological test. SAF[-] is calculated by summing up the time spent in a given neuropsychological test that intersects with a segment of time that is not DE[+] salient and dividing by the total time spent in a given neuropsychological test. The number of samples for SAF[+] and SAF[-] indicate the number of true positive and true negative recordings that contain each neuropsychological test.

References

1. Libon DJ, Swenson R, Ashendorf L, Bauer RM, Bowers D. Edith Kaplan and the Boston Process Approach. *Clin Neuropsychol* 2013; **27**(8): 1223-33.
2. Hinton G. Deep Learning-A Technology With the Potential to Transform Health Care. *JAMA* 2018; **320**(11): 1101-2.
3. Tsao CW, Vasan RS. Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *Int J Epidemiol* 2015; **44**(6): 1800-13.
4. Au R, Piers RJ, Devine S. How technology is reshaping cognitive assessment: Lessons from the Framingham Heart Study. *Neuropsychology* 2017; **31**(8): 846-61.
5. Jak AJ, Preis SR, Beiser AS, et al. Neuropsychological Criteria for Mild Cognitive Impairment and Dementia Risk in the Framingham Heart Study. *J Int Neuropsychol Soc* 2016; **22**(9): 937-43.
6. McGrath ER, Beiser AS, DeCarli C, et al. Blood pressure from mid- to late life and risk of incident dementia. *Neurology* 2017; **89**(24): 2447-54.
7. Satizabal CL, Beiser AS, Chouraki V, Chene G, Dufouil C, Seshadri S. Incidence of Dementia over Three Decades in the Framingham Heart Study. *N Engl J Med* 2016; **374**(6): 523-32.
8. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 1984; **34**(7): 939-44.
9. Wollacott AM, Xue C, Qin Q, et al. Quantifying the nativeness of antibody sequences using long short-term memory networks. *Protein Eng Des Sel* 2019; **32**(7): 347-54.
10. Gonzalez GM, Costello CR, La Tourette TR, Joyce LK, Valenzuela M. Bilingual telephone-assisted computerized speech-recognition assessment: is a voice-activated computer program a culturally and linguistically appropriate tool for screening depression in English and Spanish? *Cult Divers Ment Health* 1997; **3**(2): 93-111.
11. Kim HG, Geppert J, Quan T, Bracha Y, Lupo V, Cutts DB. Screening for postpartum depression among low-income mothers using an interactive voice response system. *Matern Child Health J* 2012; **16**(4): 921-8.
12. Munoz RF, McQuaid JR, Gonzalez GM, Dimas J, Rosales VA. Depression screening in a women's clinic: using automated Spanish- and English-language voice recognition. *J Consult Clin Psychol* 1999; **67**(4): 502-10.
13. Ozkanca Y, Ozturk MG, Ekmekci MN, Atkins DC, Demiroglu C, Ghomi RH. Depression Screening from Voice Samples of Patients Affected by Parkinson's Disease. *Digit Biomark* 2019; **3**(2): 72-82.
14. Arora S, Visanji NP, Mestre TA, et al. Investigating Voice as a Biomarker for Leucine-Rich Repeat Kinase 2-Associated Parkinson's Disease. *J Parkinsons Dis* 2018; **8**(4): 503-10.
15. Postuma RB. Voice changes in prodromal Parkinson's disease: Is a new biomarker within earshot? *Sleep Med* 2016; **19**: 148-9.
16. Tracy JM, Ozkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. *J Biomed Inform* 2020; **104**: 103362.

17. Arora S, Baghai-Ravary L, Tsanas A. Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. *J Acoust Soc Am* 2019; **145**(5): 2871.
18. Kourtis LC, Regele OB, Wright JM, Jones GB. Digital biomarkers for Alzheimer's disease: the mobile/ wearable devices opportunity. *NPJ Digit Med* 2019; **2**.
19. Gold M, Amatniek J, Carrillo MC, et al. Digital technologies as biomarkers, clinical outcomes assessment, and recruitment tools in Alzheimer's disease clinical trials. *Alzheimers Dement (N Y)* 2018; **4**: 234-42.
20. Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine* 2020; **28**: 100583.
21. Thomas JA, Burkhardt HA, Chaudhry S, 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**(3): 905-22.

Table 1

(A)

Normal cases						
Cohort	**N**	**Female**	**ApoE4+**	**Recordings**	**Age (years)**	**Mean MMSE**
Gen 1	42	28	6	75	90.9±3.0	27.5±2.0
Gen 2	238	117	40	392	73.9±7.8	28.1±2.0
Gen 3	4	1	0	7	60.4±11.5	NA
OmniGen 1	7	2	2	9	71.4±9.3	26.6±2.6
Total	291	148	48	483	76.3±9.8	27.9±2.0

(B)

MCI cases						
Cohort	**N**	**Female**	**ApoE4+**	**Recordings**	**Age (years)**	**Mean MMSE**
Gen 1	64	47	13	85	91.4±3.1	26.0±2.4
Gen 2	235	124	67	353	79.3±6.8	27.2±2.2
Gen 3	1	0	0	1	60.0±0.0	NA
OmniGen 1	6	1	1	7	73.1±6.8	25.3±2.3
OmniGen 2	1	1	1	1	74.0±0.0	NA
NOS	2	1	0	4	86.5±5.3	NA
Total	309	267	82	451	81.5±8.0	26.9±2.3

(C)

Dementia cases						
Cohort	**N**	**Female**	**ApoE4+**	**Recordings**	**Age (years)**	**Mean MMSE**
Gen 1	78	56	16	99	92.2±3.1	21.3±5.7
Gen 2	139	84	40	224	82.1±6.7	23.3±5.5
Gen 3	1	0	0	1	80.0±0.0	NA
OmniGen 1	3	1	0	4	71.5±1.7	24.8±1.5
OmniGen 2	1	1	1	1	77.0±0.0	NA
NOS	1	1	0	1	80.0±0.0	NA
Total	223	143	57	330	84.9±7.6	22.7±5.6

Figure 1

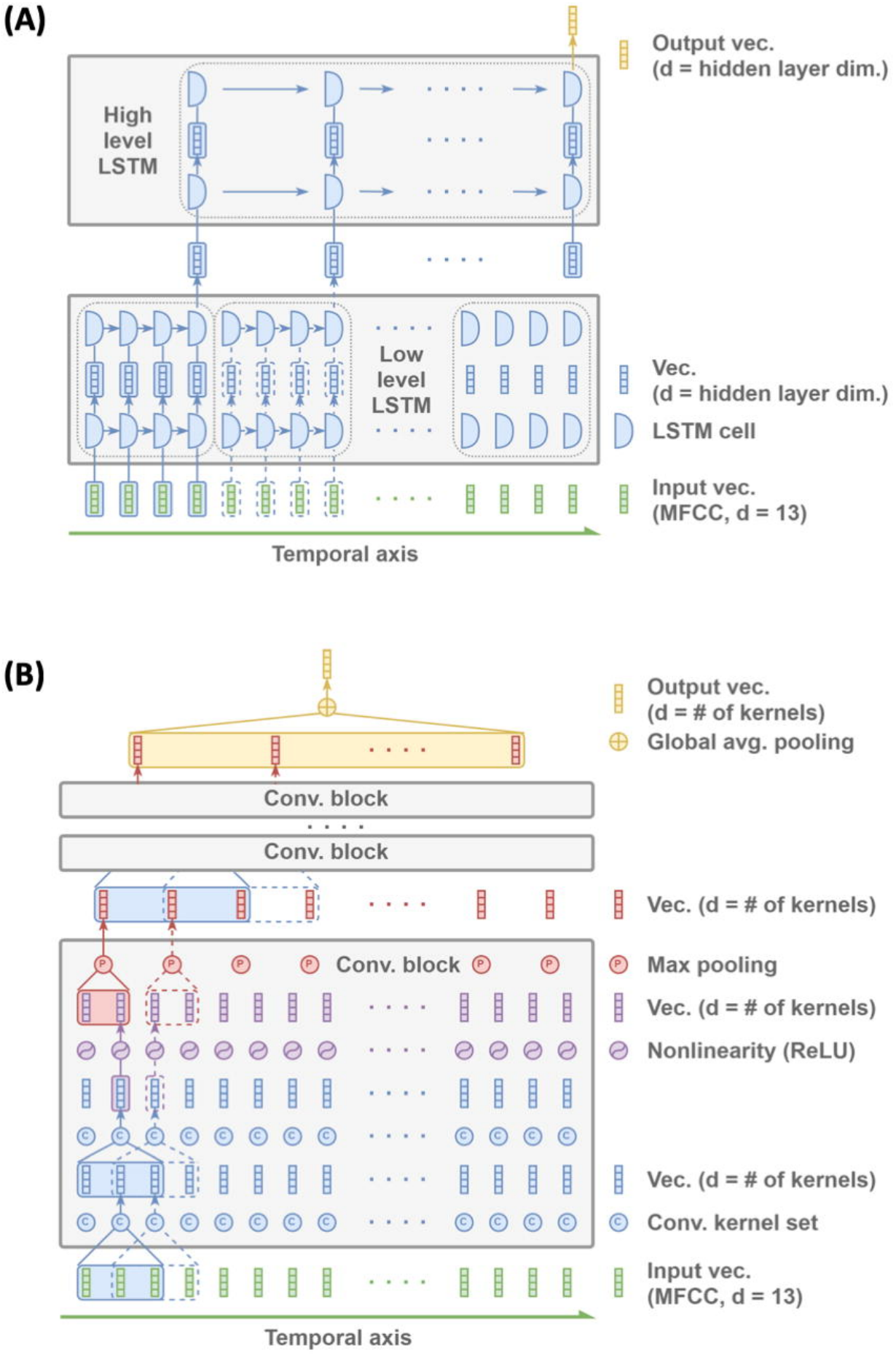
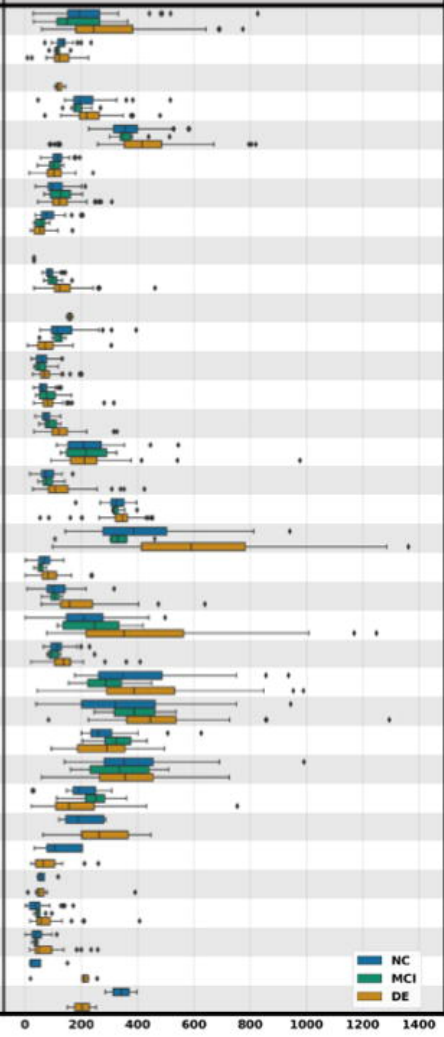

(A)
High level LSTM
Low level LSTM
Output vec.
(d = hidden layer dim.)
Vec.
(d = hidden layer dim.)
LSTM cell
Input vec.
(MFCC, d = 13)
Temporal axis
(B)
Output vec.
(d = # of kernels)
Global avg. pooling
Conv. block
Conv. block
Vec. (d = # of kernels)
Conv. block
Max pooling
Vec. (d = # of kernels)
Nonlinearity (ReLU)
Vec. (d = # of kernels)
Conv. kernel set
Input vec.
(MFCC, d = 13)
Temporal axis

Figure 2



Neuropsychological Test	NC #	MCI #	NonDE #	NC_MCI	MCI_DE	NC_DE	NonDE_DE
Demographics	35	9	71	ns	*	ns	*
Logical Memory Immediate Recall	43	10	78	ns	ns	ns	ns
Logical Memory Immediate Recall (†)	0	0	3	n/a	n/a	n/a	n/a
Visual Reproductions Immediate Recall	43	10	70	ns	*	ns	ns
Verbal Paired Associates Learning	44	10	72	ns	ns	ns	*
Digit Span Forward	44	10	75	ns	ns	ns	ns
Digit Span Backward	43	10	75	ns	ns	*	*
Logical Memory Delayed Recall	43	10	71	*	ns	****	****
Logical Memory Delayed Recall (†)	0	0	2	n/a	n/a	n/a	n/a
Logical Memory Multiple Choice	43	9	71	ns	ns	****	****
Logical Memory Multiple Choice (†)	0	0	2	n/a	n/a	n/a	n/a
Visual Reproductions Delayed Recall	41	9	61	ns	*	****	****
Visual Reproductions Multiple Choice	41	9	61	ns	ns	*	*
Verbal Paired Associates Recall	43	9	58	ns	ns	**	**
Verbal Paired Associates Recognition	42	9	57	ns	*	****	****
Similarities	42	9	73	ns	ns	ns	ns
Command Clock	41	9	70	ns	ns	***	****
Verbal Fluency (FAS)	41	9	72	ns	ns	ns	ns
Boston Naming Test	41	9	70	ns	**	****	****
Copy Clock	36	9	63	ns	*	*	**
Trails A	38	9	59	ns	*	****	****
Trails B	38	9	52	ns	*	****	****
WRAT-3 Reading	37	9	47	ns	ns	ns	*
Hooper Visual Organization Test	38	9	58	ns	*	ns	ns
Block Design (WAIS)	24	3	37	ns	ns	*	*
Finger Tapping	20	5	15	ns	ns	ns	ns
Information (WAIS-R)	26	4	49	ns	ns	ns	ns
Cookie Theft	20	4	27	ns	ns	ns	ns
Digit Symbol Coding	5	0	14	n/a	n/a	ns	ns
Digit Symbol Learning	5	0	11	n/a	n/a	ns	ns
Digit Symbol Recall	5	0	8	n/a	n/a	ns	ns
Clock Drawing Number Placement	35	9	38	ns	ns	*	**
Clock Drawing Time Setting	34	8	37	ns	ns	**	**
Math Fluency	4	0	6	n/a	n/a	*	*
Balance Physical Function Test	2	0	2	n/a	n/a	ns	ns

Figure 3

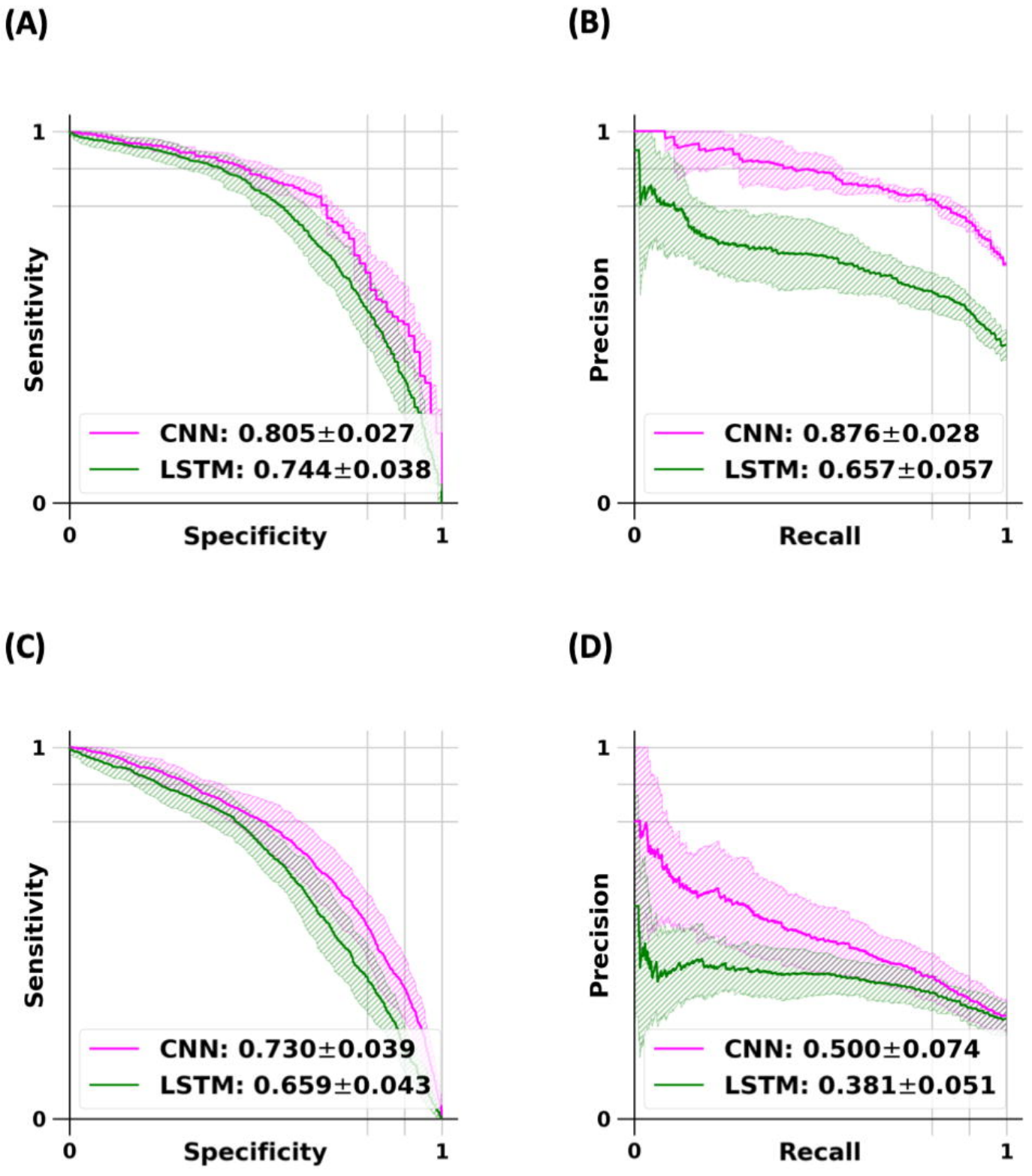

(A)
1
0
Sensitivity
0
Specificity
1
CNN: 0.805±0.027
LSTM: 0.744±0.038
(B)
1
0
Precision
0
Recall
1
CNN: 0.876±0.028
LSTM: 0.657±0.057
(C)
1
0
Sensitivity
0
Specificity
1
CNN: 0.730±0.039
LSTM: 0.659±0.043
(D)
1
0
Precision
0
Recall
1
CNN: 0.500±0.074
LSTM: 0.381±0.051

Figure 4

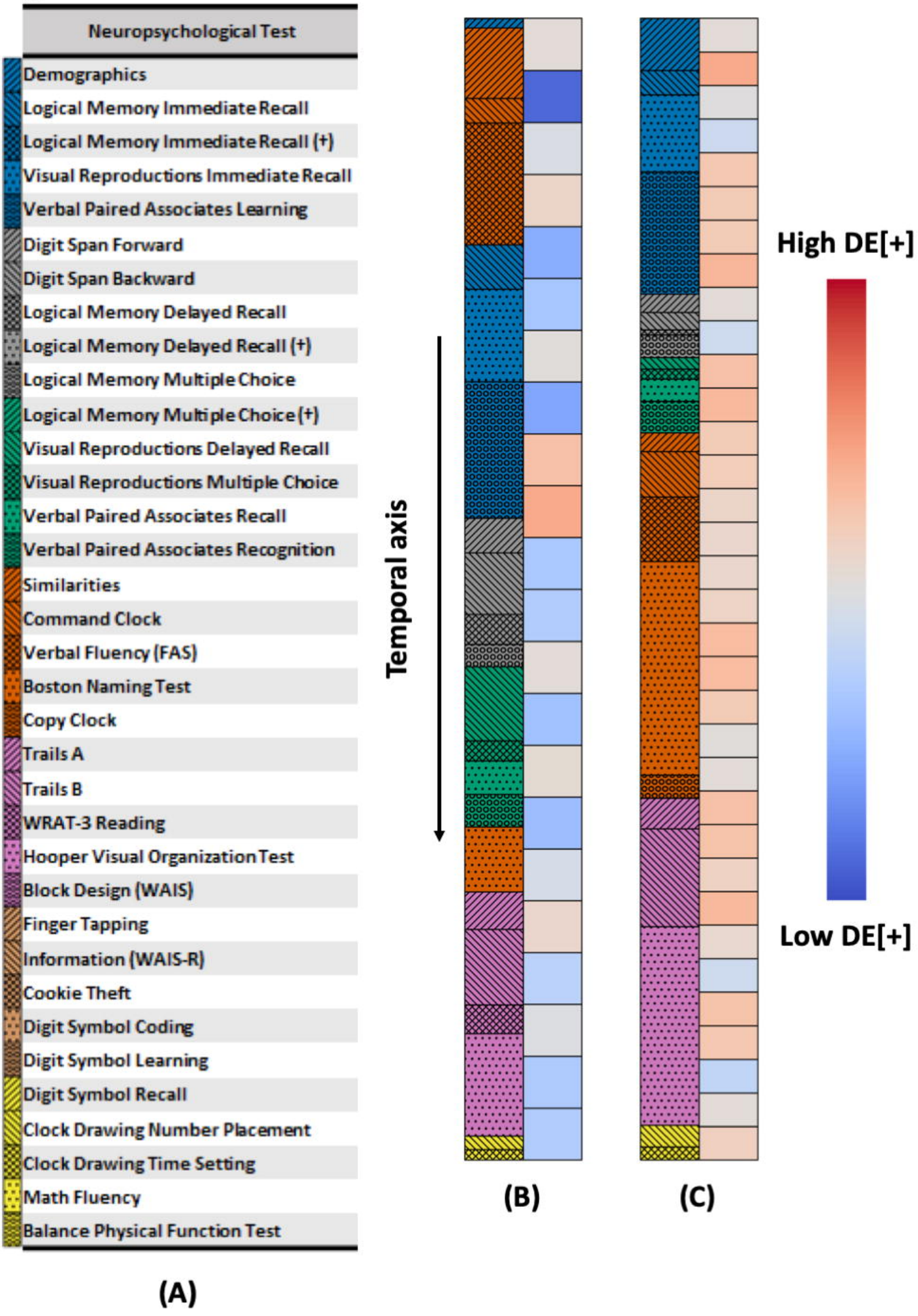

Neuropsychological Test
Demographics
Logical Memory Immediate Recall
Logical Memory Immediate Recall (+)
Visual Reproductions Immediate Recall
Verbal Paired Associates Learning
Digit Span Forward
Digit Span Backward
Logical Memory Delayed Recall
Logical Memory Delayed Recall (+)
Logical Memory Multiple Choice
Logical Memory Multiple Choice (+)
Visual Reproductions Delayed Recall
Visual Reproductions Multiple Choice
Verbal Paired Associates Recall
Verbal Paired Associates Recognition
Similarities
Command Clock
Verbal Fluency (FAS)
Boston Naming Test
Copy Clock
Trails A
Trails B
WRAT-3 Reading
Hooper Visual Organization Test
Block Design (WAIS)
Finger Tapping
Information (WAIS-R)
Cookie Theft
Digit Symbol Coding
Digit Symbol Learning
Digit Symbol Recall
Clock Drawing Number Placement
Clock Drawing Time Setting
Math Fluency
Balance Physical Function Test
(A)
Temporal axis
(B)
(C)
High DE[+]
Low DE[+]

Table 2

(A)

Normal vs. Demented classification		
Model	**LSTM**	**CNN**
Accuracy	0.680±0.032	**0.740±0.033**
Sensitivity	0.719±0.112	**0.735±0.094**
Specificity	0.652±0.089	**0.750±0.083**
Precision	0.595±0.057	**0.844±0.034**
F1 Score	0.645±0.058	**0.780±0.048**
MCC	0.369±0.069	**0.477±0.026**
Precision-Recall AUC	0.657±0.057	**0.876±0.028**
Sensitivity-Specificity AUC	0.744±0.038	**0.805±0.027**

(B)

Non-demented vs. Demented classification		
Model	**LSTM**	**CNN**
Accuracy	0.701±0.057	**0.735±0.046**
Sensitivity	0.245±0.161	**0.443±0.113**
Specificity	**0.856±0.105**	0.840±0.076
Precision	0.383±0.103	**0.510±0.104**
F1 Score	0.271±0.129	**0.459±0.072**
MCC	0.117±0.078	**0.299±0.083**
Precision-Recall AUC	0.381±0.051	**0.500±0.074**
Sensitivity-Specificity AUC	0.659±0.043	**0.730±0.039**

Table 3

Neuropsychological test	SAF[+]	SAF[-]	SAF[+] samples	SAF[-] samples
Verbal paired associates recognition	0.88±0.26	0.32±0.42	37	32
Verbal paired associates learning	0.83±0.22	0.23±0.22	51	33
Boston naming test	0.82±0.18	0.30±0.29	47	31
Digit span forward	0.78±0.37	0.16±0.33	53	33
Wrat-3 reading	0.76±0.36	0.57±0.43	30	27
Information (WAIS-R)	0.73±0.28	0.36±0.30	36	16
Verbal paired associates recall	0.72±0.41	0.33±0.41	38	32
Finger tapping	0.71±0.33	0.47±0.33	11	15
Digit span backward	0.70±0.42	0.36±0.43	53	32
Demographics	0.69±0.33	0.22±0.35	55	25
Trails B	0.68±0.30	0.65±0.35	34	28
Logical memory immediate recall	0.66±0.42	0.72±0.40	56	32
Verbal fluency (FAS)	0.66±0.34	0.39±0.27	49	31
Similarities	0.66±0.38	0.56±0.34	50	32
Clock drawing time setting	0.66±0.48	0.48±0.51	24	25
Logical memory delayed recall	0.65±0.44	0.74±0.36	50	32
Cookie theft	0.64±0.41	0.65±0.39	19	14
Block design (WAIS)	0.64±0.27	0.70±0.33	26	14
Clock drawing number placement	0.63±0.45	0.49±0.49	26	26
Unlabeled	0.61±0.50	0.25±0.44	31	20
Visual reproductions multiple choice	0.60±0.47	0.44±0.48	41	30
Hooper visual organization test	0.58±0.29	0.48±0.26	39	28
Logical memory multiple choice	0.57±0.42	0.87±0.28	50	32
Miscellaneous	0.57±0.39	0.47±0.45	36	15
Visual reproductions immediate recall	0.55±0.35	0.85±0.21	48	32
Trails A	0.52±0.41	0.51±0.44	40	28
Visual reproductions delayed recall	0.52±0.47	0.77±0.33	40	30
Copy clock	0.44±0.45	0.36±0.47	41	26
Command clock	0.39±0.45	0.76±0.37	47	31

Table S1

Non-demented cases						
Cohort	**N**	**Female**	**ApoE4+**	**Recordings**	**Age (years)**	**Mean MMSE**
Gen 1	89	60	15	160	91.2±3.1	26.7±2.3
Gen 2	398	201	89	745	76.5±7.8	27.6±2.1
Gen 3	5	1	0	8	60.4±10.6	NA
OmniGen 1	12	3	3	16	72.2±8.1	26.0±2.5
OmniGen 2	1	1	1	1	74.0±0.0	NA
NOS	2	1	0	4	86.5±5.3	NA
Total	507	267	108	934	78.8±9.3	27.4±2.2

Table S2

Test	SAF[+]	SAF[-]	SAF[+] samples	SAF[-] samples
Balance Physical Function Test	1.00±0.00	0.63±0.30	2	2
Logical Memory Delayed Recall (alternate prompt)	1.00±0.00		1	0
Digit Symbol Learning	0.77±0.41	0.81±0.33	6	3
Math Fluency	0.74±0.34		4	0
Digit Symbol Recall	0.74±0.46	0.40±0.53	3	3
Logical Memory Multiple Choice (alternate prompt)	0.64±0.00		1	0
Logical Memory Immediate Recall (alternate prompt)	0.46±0.18		2	0
Digit Symbol Coding	0.43±0.43	1.00±0.00	9	3