

Automated Detection of Mild Cognitive Impairment and Dementia from

Voice Recordings: a Natural Language Processing Approach*

Samad Amini^a, Boran Hao^a, Lifu Zhang^a, Mengting Song^a, Aman Gupta^a, Cody Karjadi^c, Vijaya B.

Kolachalama^{b,d,e}, Rhoda Au^{f,c}, Ioannis Ch. Paschalidis^{a,d,g}

^aDepartment of Electrical & Computer Engineering, Division of Systems Engineering, and Department of Biomedical Engineering, Boston University

^bDepartment of Medicine, Boston University School of Medicine

^cFramingham Heart Study, Boston University

^dFaculty of Computing & Data Sciences, Boston University

^eDepartment of Computer Science, Boston University

^fDepartments of Anatomy & Neurobiology, Neurology, and Epidemiology, Boston University School of Medicine and School of

^gPublic Health

^gCorresponding author: Ioannis Ch. Paschalidis, yannisp@bu.edu, 8 St. Mary's St Boston, MA 02215

Abstract

INTRODUCTION: Automated computational assessment of neuropsychological tests would enable wide-

spread, cost-effective screening for dementia.

METHODS: A novel natural language processing approach is developed and validated to identify different

stages of dementia based on automated transcription of digital voice recordings of subjects' neuropsychological tests

conducted by the Framingham Heart Study ($n = 1,084$). Transcribed sentences from the test

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1002/alz.12721](https://doi.org/10.1002/alz.12721).

This article is protected by copyright. All rights reserved.

20 were encoded into quantitative data and several models were trained and tested using this
data and the

21 participants' demographic characteristics.

22 **RESULTS:** Average Area Under the Curve (AUC) on the held-out test data reached 92.6%,
88.0%, and

23 74.4% for differentiating Normal cognition from Dementia, Normal or Mild Cognitive
Impairment (MCI)

24 from Dementia, and Normal from MCI, respectively.

25 **DISCUSSION:** The proposed approach offers a fully automated identification of MCI and
dementia based

26 on a recorded neuropsychological test, providing an opportunity to develop a remote screening tool
that could

27 be easily adapted to any language.

28 **Keywords:** Alzheimer's disease, Cognitive impairment, Natural Language Processing,
Neuropsychological

29 tests, Framingham Heart Study.

Disclosures: Rhoda Au is a scientific advisor to Signant Health, consultant to Biogen, and has received support from Pfizer and GlaxoSmithKline. She has also been supported through grants by the American Heart Association, the Alzheimer's Drug Discovery Foundation, and Gates Ventures, and has received other support from Eisai, Gates Ventures and the High Lantern Group. Vijaya B. Kolachalama has received support from Johnson & Johnson (through the Boston University Lung Cancer Alliance), the NIH under grants R21 CA253498 and 5U24DK115255, the American Heart Association under grant 20SFRN35460031, and has received speaker fees from MassMutual. Both R. Au and V. B. Kolachalama state no conflicts of interest with the present work. There is no declaration from other authors.

*The research was partially supported by the NSF under grants DMS-1664644, CNS-1645681, and IIS-1914792, by the ONR under grant N00014-19-1-2571, by the DOE under grants DE-AR-0001282 and DE-EE0009696, by the NIH under grants R01 GM135930, UL54 TR004130, R01-HL159620 and R21-CA253498, by the Framingham Heart Study's National Heart, Lung, and Blood Institute contract (N01-HC-25195; HHSN268201500001I), by the NIH National Institute on Aging (AG008122, AG016495, AG033040, AG054156, AG049810, AG062109, AG068753), by the Alzheimer's Association under grant AARG-NTF-20-643020, by Pfizer, by a Toffler scholarship in neuroscience from the Karen Toffler Charitable Trust, by the American Heart Association's Strategically Focused Research Networks grant (20SFRN35460031), and by Boston University.

Preprint submitted to Journal of Alzheimer's & Dementia March 20, 2022

This article is protected by copyright. All rights reserved.

1. Introduction

In recent years, there has been a growing interest in developing novel technological approaches to aid in

the clinical diagnosis and assessment of *Alzheimer's disease (AD) and Related Dementias (ARD)*. In the

U.S., more than 5 million individuals are living with AD – the most common form of dementia – with AD

deaths increasing by 146% between 2000 and 2018 [1].

AD pathology can be assessed using (a) biomarkers such as amyloid- β plaques and neurofibrillary tangles

of tau protein detected in cerebrospinal fluid and observed through Positron Emission Tomography (PET)

and (b) evidence of neurodegeneration assessed through Magnetic Resonance Imaging (MRI) [2, 3, 4]. These

testing modalities imply a well-resourced setting and yet do not necessarily determine cognitive decline, since,

in some cases, AD brain pathology may not translate into clinical expression. Triggered by patient history

and in conjunction with a clinical examination, a *neuropsychological (NP)* exam, conducted through an in-

person interview, is currently the primary method for assessing cognitive decline, even at early stages. The

Framingham Heart Study (FHS) NP tests take 45–90 minutes and cover all major cognitive domains mostly

through connected speech [5], which is a cognitively intense activity [6, 7]. FHS has been recording its NP

This article is protected by copyright. All rights reserved.

15 tests since 2005.

16 Studies have shown that speech can be a strong predictor of cognitive impairment in early stages [8, 9].

17 Machine learning algorithms have been developed to build diagnostic models using vocal and lexical features

18 extracted from voice recordings [10, 11, 12, 13]. Vocal features were extracted in [14] to develop a classifier

19 to predict dementia among 64 subjects. Fraser et al. [15] have focused on linguistic as well as vocal features;

20 achieving accuracy of 81.9% in classifying dementia versus control. Mild Cognitive Impairment (MCI) was

21 also identified from healthy cases (AUC = 80%) by applying the support vector machine on features extracted

22 from manual transcriptions of voice recordings [16]. Using the FHS battery of tests, Lin et al. presented a

23 voice-based predictor to identify incidents of dementia with an AUC of 81% [17]. In addition, language and

24 voice features of 170 participants from FHS were analyzed to predict cognitive impairment with an AUC of

25 94% [4].

26 However, most studies relied on both manual transcription and handcrafted features of the voice data.

27 Manual transcription is a lengthy and expensive procedure which might hinder its implementation in a large-

28 scale setting. Almost all the findings were limited from the use of relatively small sample sizes and disparities

29 observed in the clinical population. In particular, in the context of speech assessment and machine learning,

30 very few researchers included both MCI and dementia cases in their studies [5, 18]. Given the fact that

31 cognitive decline can take several years to evolve from MCI to more severe stages of dementia, automated

32 detection of MCI is crucial to allow effective intervention in early stages. Furthermore, considering the

1 shift towards virtual visits in response to the ongoing health care issues, such as the COVID-19 pandemic,

2 it is highly desirable to develop easily deployable, cost-effective, automated and accurate MCI/dementia

3 assessment procedures with the potential to drastically increase the pool of candidates for ADRD clinical

4 trials, possibly accelerating the search for effective treatments. To that end, so called digital biomarkers have

5 an important role to play [19].

6 We sought to design an automated diagnostic tool to detect MCI and dementia based on voice recordings.

7 The proposed method takes either an audio file or a transcript of the NP interview as input and predicts the

8 likelihood of cognitive impairment using machine learning techniques, leveraging *Natural Language Processing*

9 (*NLP*). We rigorously evaluated our approach on automated transcriptions and achieved promising prediction

10 performance. Recently, studies have shown reliable performance of NLP algorithms in health care (see [20])

This article is protected by copyright. All rights reserved.

11 for a review). Moreover, NLP has made it possible to automatically extract information from
unstructured

12 data (text, audio, video) in medical records [21, 22, 23]. Our study uses a variety of NLP methods,
including

13 speech recognition, speech diarization, and a transformer-based sentence encoder.

14 **2. Materials and Methods**

15 *2.1. Clinical setting and data sources*

16 The digital recordings used in this study were collected from the NP examination administered by
FHS,

17 the longest ongoing longitudinal, transgenerational cohort study of chronic disease [24]. The NP
tests include

18 sub-tests that assess naming and language, visuoperceptual skills, premorbid intelligence, abstract
reasoning,

19 attention, verbal memory (logical memory immediate and delayed recall), learning (paired-
associate memory

20 immediate and delayed recall), and visual memory (visual reproductions immediate and delayed
recall) [25,

21 26]. Additional information such as sex, age, education, and presence of Apolipoprotein E
(ApoE) genes

22 is available as well. The current study includes recordings of subjects evaluated by trained
examiners from

23 2005 to 2015. The participants' cognitive status was determined by the FHS dementia
diagnostic review

24 panel [27]. Dementia diagnosis for those showing signs of cognitive decline was reached by
consensus of at

25 least one neurologist and one neuropsychologist based on neurology exams, FHS study and
external medical

26 records, and brain imaging (the diagnostic procedure is outlined in [25, 28]). All participants have
provided

27 written informed consent and study protocols and consent forms were approved by the Boston
University

28 Medical Campus Institutional Review Board.

29 *2.2. Data preparation*

30 The original dataset contains information about 1,084 participants, including audio files stored in
.wav

31 format (8Khz sampling rate), age, sex, education, ApoE, and dementia diagnosis. A total of
133 of these

1 recordings have been manually transcribed, where each transcript is separated into 8 sub-tests
(details of

2 these sub-tests can be found in the supplementary). In order to prepare the material for the
proposed method,

3 all the audio files were transcribed via the Google Speech tool [29]. It has been shown that
Google Speech

4 achieved a 9% word error rate, outperforming other well-known speech recognition systems such as
Sphinx-4

5 and Microsoft [30]. Given the raw text files generated by Google Speech, each transcript was
diarized and

6 sentences labeled using NLP to indicate the specific sub-test they belong. It should be noted that
the Google

7 Speech tool can separate the speakers in an audio file but it is unable to identify the label of the speakers,

8 in our case, examiner and participant. For diarization, we fine-tuned an ALBERT-xlarge model [31] to

9 separate sentences of the participants (P) and the examiners (T) in each transcript, generating a collection of

10 unpunctuated and uncased words. To that end, we used a Named Entity Recognition (NER)-type approach.

11 The ALBERT model used context to assign to each word one of the 4 labels: B-T, I-T, B-P and I-P, where

12 'B-' means beginning and 'I-' means inside. Then, a complete P sentence can be detected using a 'B-P' word

13 and the following consecutive 'I-P' words. Splitting the 133 manually-transcribed recordings into a training

14 and test set according to 80:20 ratio, we trained and tested the ALBERT model, obtaining an Exact F1-score

15 of 70.2% on the test set. The Exact F1-score is a strict metric in NER tasks — even if we miss one word in

16 a target sentence, which does not affect the semantic, we still regard this target as misclassified.

17 We further fine-tuned a BERT-based model [32] to predict the 8 different sub-tests for each P/T sentence

18 detected in the previous step. To improve the prediction accuracy and consistency, we used the target

19 sentence S_0 together with the prior and following 8 consecutive sentences (“[CLS] S_{-8} [SEP]...[SEP] S_0

20 [SEP]...[SEP] S_8 [SEP]”) as the BERT input, and the [CLS] embedding was used as the sub-test predictor

21 for S_0 . If certain neighbor sentences do not exist, such positions were ignored. Using the manual transcription

22 and the ground truth labels of the sub-tests to train the BERT, the accuracy of classifying sentences into

23 the 8 sub-test types reached 96.2%. This sub-test classification step was done separately and the accuracy

24 was obtained without using the output of the diarization step because the diarization output lacks the sub-

25 test ground truth label. Figure 1 demonstrates the mechanism of diarization and sub-test labeling in the

26 pre-processing procedure. In this analysis, the sentences can consist of an arbitrary number of words with

27 any structure. Some sentences can be as short as a single word or as long as a paragraph. Both BERT

28 and ALBERT are state-of-the-art methods in NLP. We considered both models for the pre-processing tasks.

29 However, the ALBERT model has fewer parameters compared to a corresponding BERT model (e.g., 60

30 million parameters vs. 334 million parameters), resulting in lower memory usage and lower computational

31 cost. Since the performance of both models for the diarization task was the same, the ALBERT model was

32 selected as our method of choice. In the sub-test classification task, the BERT model outperformed the

33 ALBERT model with a margin of about 1% accuracy on 133 manually transcribed recordings. Therefore,

1 the BERT model was used to identify the sub-tests.

2 2.3. Statistical analysis

3 The dataset includes participants who were labeled as normal, with MCI, or dementia (mild,
4 moderate,
5 severe), respectively. Various basic characteristics are reported in Table 1. We report self-
6 reported gender,
7 education status, age statistics (mean \pm standard deviation for each cohort), dementia
8 diagnosis severity,
9 and the type of ApoE (E2/E3/E4) genes for both copies of the gene. We also report the p -value
10 for each
11 variable associated with the null hypothesis that the two cohorts have the same distribution of the
12 variable.
13 Hence, a low p -value indicates that the null hypothesis should be rejected, suggesting that the
14 distribution
15 of the feature is different in each cohort. For age, we employed the Kolmogorov-Smirnov (K-S)
16 test [33],
17 whereas we used the Chi-square test for the categorical features (sex, education, and ApoE) [34].

11 2.4. Development of the model

12 The NLP-based detection system can be formulated as a classification task. We have investigated
13 three
14 binary classification tasks: (I) dementia detection (normal versus dementia), (II) non-dementia
15 detection
16 (normal/MCI versus dementia), and (III) MCI detection (normal versus MCI). Due to the
17 limited data
18 both in terms of sample size and composition – we used a transfer learning approach to
19 capture relevant

16 characteristics of the participants' sentences, subsequently converting the text data into a machine readable

17 format. Transfer learning is a widely used technique in NLP applications that addresses the problem of

18 training a classifier when a large, complete training dataset is not available [35].

19 After the pre-processing step outlined in Section 2.2, the Universal Sentence Encoder (USE) was employed

20 to encode the P sentences into an embedding vector. NP assessment is typically based on the participants'

21 responses [16, 36]. The USE, which is a neural network based on the transformer architecture and attention

22 mechanism, has provided state-of-the-art results on a different range of tasks [37]. This encoder is pre-trained

23 on a variety of sources like Wikipedia, web news, web question-answer pages, and discussion forums [38].

24 Once the quantitative data (the encoded sentences) were generated, our system computed the likelihood

25 of whether an individual is cognitively impaired using machine learning techniques such as the Multilayer

26 Perceptron (MLP) and logistic regression. Figure 2 provides an overview of the proposed system.

27 The USE was selected to encode the sentences because it takes an input text with an arbitrary length

28 whereas the length of input text in other methods is often limited (e.g., the BERT model takes as input a

29 sequence of no more than 512 tokens, roughly equal to 400 words). Although the USE can process text data

30 of any length, feeding the entire transcript into the USE would result in a poor downstream classification

31 performance, mainly because USE generates a fixed 512-dimensional vector for any input. Instead, we

32 implemented two different approaches that enable us to extract more information from each transcript.

1. *Random Sampling (RS) method*: We constructed a paragraph by randomly selecting 25 P sentences

from each interview, given that each transcript contains at least 25 P sentences. We repeated the

random sampling for each interview until 30 paragraphs were collected. These 30 paragraphs are

different from each other even if a transcript contains only 25 P sentences due to random order of the

sentences. Then, the USE takes each paragraph as input and generates a 512-dimensional sentence

embedding.

2. *Sub-test Sampling (STS) method*: The STS method exploits the sub-test labels of the P sentences.

This method groups the P sentences of the same sub-test together, compiling eight paragraphs from

each interview (one for each sub-test). By passing these paragraphs to the USE, eight 512-dimensional

10 vectors are generated for each interview. If any sub-test is missing throughout the interview, we fill the

11 corresponding vector with zeroes.

12 After obtaining a vector representing a paragraph created by the above method, we proceeded with

13 the classification task. We processed the embedding vectors in three steps: (i) feature
selection of the

14 embedding vectors, to improve downstream classification performance, (ii) generating cognitive
scores from

15 the resulting lower dimensional vectors, and (iii) training a classifier by a combination of the
cognitive

16 scores and demographic information to arrive at the final prediction. Given the training and
test dataset,

17 we performed logistic regression-based recursive feature elimination (e.g., as in [39]) on the
training data to

18 remove the weakest feature until the embedding vector is of size 50, resulting in reducing the 512-
dimensional

19 vectors to 50-dimensional vectors.

20 To generate the cognitive scores, the RS method trains an MLP model on the low-dimensional
embedding

21 vectors. The model treats the extracted paragraphs of an interview as independent examples with
the same

22 label as the corresponding interview label. In the second method (STS), we trained 8 MLP models
separately

23 for each sub-test using the corresponding low-dimensional embedding vectors. All the MLP
models used in

24 this step were networks with an input layer, an output layer, and a hidden layer with 25
hidden nodes.

25 Therefore, the first method generated 30 cognitive scores (one per paragraph) for each person using
only one

26 predictive model, whereas the other method generated 8 cognitive scores associated with each sub-
test, using

27 8 independent MLP models. In the RS method, we then trained a logistic regression model using
the average

28 of the 30 cognitive scores in addition to demographic information.

29 In the STS method, however, a subset of the sub-tests were selected to be fed into the logistic
regression

30 model along with the demographics. To select the important sub-tests, we calculated the
increase in the

31 model's prediction error using the validation data after removing the sub-test during the
training of the

32 logistic regression model. Comparing the model's performance error, we identified a relative
ranking of the

1 sub-tests (see Table 1 of the Supplement for details and naming conventions). These rankings are
different for

2 each one of the three classification tasks. Specifically, the sub-tests ranked from most to least
important were

3 as follows: (i) [BNT, WMS, FAS, OTHER, CDT, WAIS-R, WAIS, DEMO] for Task I, (ii) [WMS,
BNT, CDT,

4 FAS, OTHER, WAIS-R, WAIS, DEMO], for Task II, and [WAIS-R, BNT, WMS, DEMO,
OTHER, WAIS,

5 CDT, FAS] for Task III, respectively. To select the best set of sub-tests, we evaluated the
performance of the

6 n most important sub-tests on the validation dataset. Therefore, out of 8 sub-tests, the STS
method ended

7 up achieving the best AUC performance on [BNT, WMS, FAS, OTHER, CDT, WAIS-R],
[WMS, BNT], and

8 [WAIS-R, BNT, WMS, DEMO, OTHER] in Task I, II, and III, respectively (see Table 1 of the Supplement

9 for an explanation of the sub-tests). In addition to these NLP-based methods, we developed an ensemble

10 model that combines multiple other models in the prediction process. For instance, one can train a logistic

11 regression model that combines the RS and/or STS methods with different variables such as demographics or

12 the presence of ApoE genes. The entire prediction procedure was implemented using the python deep-learning

13 *Keras* library with a *Tensorflow* backend.

14 2.5. Validation and performance metrics

15 The data were randomly split into 10 folds using stratified k -fold cross validation. A model was trained

16 on the 9 folds and tested on the 10th – test – fold. The training process was repeated 10 times for all ten

17 models to ensure the accuracy of the results. Performance metrics consisted of classification accuracy (Acc),

18 the Area Under the Receiver Operating Characteristic Curve (AUC), sensitivity, and specificity [40]. AUC is

19 a useful measure that indicates the probability of the classifier ranking a randomly selected positive sample

20 higher than a randomly selected negative sample. Sensitivity and specificity show how accurately positive

21 subjects and negative subjects are classified.

22 3. Results

23 The results of various methods based on automated transcriptions are summarized in Table 2. In this table,

24 but also for the remainder of this paper, the “+” symbol between groups of variables/methods represents an

25 ensemble model that uses both of these groups/methods to make a prediction.

26 In the first set of three rows of Table 2 (one per Task), we report the performance of the STS method

27 that utilizes the sub-test information of the NP test. The second set of three rows reports the performance

28 of the ensemble model that combines the STS method and demographic information (sex and age), which

29 leads to the 2nd highest AUC among all methods for Tasks I and II. The third set of three rows corresponds

30 to the RS method. The fourth set of three rows reports the performance of the ensemble model using the

31 RS method and demographic variables. The fifth to seventh set of three rows report the performance of the

1 baseline models using different combinations of demographic information, ApoE, and education. The 8th

2 set of three rows corresponds to an ensemble model using both methods along with age and sex. The 9th

3 set of three rows corresponds to the ensemble model that uses all of the sub-tests and demographics (age

4 and sex). In order to compare our approach with a well-established cognitive assessment tool, we considered

5 the Mini-Mental State Examination (MMSE) [41] score in a logistic regression model that performs binary

6 classification for each Task; the corresponding results are reported on the 10th set of three rows. The final

7 row is specifically for MCI detection considering STS, demographics, and ApoE, leading to the best AUC

8 among all methods for this task. In all these models, education, sex, and ApoE features, were encoded using

9 one-hot encoding, i.e., creating a binary variable for each category.

10 We observe that the ensemble model STS+RS+Dem. achieves the best AUC for Tasks I and II, equal

11 to $92.6\% \pm 3.3\%$ and $88.0\% \pm 2.2\%$, respectively. The 2nd best AUC for these tasks is achieved by the

12 STS+Dem. method, equal to $91.2\% \pm 4.1\%$ and $87.1\% \pm 4.2\%$, respectively. For Task 3, the best AUC

13 is achieved by combining STS+Dem.+ApoE, reaching $74.4\% \pm 4.4\%$, with STS+RS+Dem. coming close

14 behind (AUC: $74.1\% \pm 4.4\%$).

15 For Task I, all methods that use our NLP approach (either through the STS or the RS method), achieve

16 an AUC above 88.5%, whereas the best AUC without using NLP does not exceed 78.1%; the AUC difference

17 between the best NLP method over the best non-NLP method is 14.5%. Similarly, for Task II, NLP-based

18 methods exceed 83.7% in AUC, whereas non NLP-based methods reach up to 72.3%; the AUC difference of

19 the best NLP method over the best non-NLP method is 15.7%. Finally, for Task III, NLP-based methods

20 exceed 67.8% in AUC, whereas the non NLP methods reach 67.2% in AUC; however, the AUC
difference

21 of the best NLP method over the best non-NLP method is 7.2%. Overall, NLP methods result
into a very

22 significant boost in performance.

23 It is also worth noting that our proposed methods significantly outperform well-established
baseline

24 assessment tests based on the MMSE. As an alternative to the MMSE, more recent work has
shown that

25 it can be replaced by other tests (such as the Short Cognitive Performance Test (SKT) [18])
which achieve

26 almost identical performance.

27 A visual summary of the AUC results for each task is provided by the box plots in Figure 3a-3c.
Figure 3d

28 plots the Receiver Operating Characteristic (ROC) of the STS+RS+Dem. method, which
performs the

29 best for Tasks I and II. Figure 4 reports the coefficients of the features of the logistic regression
using the

30 STS+RS+Dem. method. The coefficients are comparable as the scores of the selected sub-tests, RS
features,

31 and age were normalized by subtracting the mean and dividing by the standard deviation. It
can be seen

32 that the RS score along with other sub-tests are contributing more to the decision than the
demographic

33 features such as age and sex.

Moreover, and to assess the impact of automated transcription on the classifications tasks, we computed

the average AUC derived from the proposed methods using manual transcriptions. Table 2 of the Supplement

reports the performance metrics of different methods on the 133 manual transcripts available. For instance,

the STS+RS+Dem. model outperformed the baseline Dem. model (Task I: average AUC of 96.1% versus

63.9%, Task II: average AUC of 94% versus 70%). The cohort associated with the manual transcripts

consisted of 45 Normal, 11 MCI, and 77 Dementia participants according to the diagnosis score given in

Table 1. Unfortunately, there exist few transcripts of the MCI class, preventing us from performing Task III

using manual transcriptions.

4. Discussion

Our fully automated system demonstrates a strong predictive power to detect cognitive impairment based

on digital voice recordings of the NP test. Due to its automatic screening ability, after prospective validation,

the proposed system can support clinicians by aiding accurate diagnosis of dementia and MCI, making it

suitable for large-scale screening of cognitive impairment. Widely accessible cognitive decline assessment

14 is not widely available even in the U.S., let alone other, less developed, countries. Therefore, our system

15 can form the basis of a diagnostic tool that is economical, particularly for less well-resourced regions and

16 for segments of the population in developed countries with insufficient access to these types of health care

17 services (e.g., rural areas, lower income individuals, underrepresented groups, etc.).

18 Another characteristic of our study is that it relies on semantic features, enabling us to transfer the

19 entire pipeline to other languages given the existence of transcription tools from any language to English

20 and/or powerful NLP models in different languages [42, 43]. At the same time, an end-to-end learning from

21 acoustic features like the Mel-frequency cepstral coefficients (MFCCs) suffers from task independence and

22 requires more resources especially in long audio files [44, 45]. We note that the performance drop due to the

23 automated transcription is rather modest, 3.5% in AUC for Task I and 6.0% for Task II, when using the

24 STS+RS+Dem. method. For comparison purposes, a recent work by Xue et. al [46] also used subjects from

25 the FHS. Using the acoustic features and deep learning methods, they achieved an AUC of 80% and 75% for

26 Task I and II, respectively. Thus, the adverse affect of including MCI cases on the classifier performance can

27 be significantly mitigated through the proposed NLP-based route.

28 Since the proposed method shows promising performance on automated transcription, a remote diagnosis

29 can be contemplated based on interviews conducted through video/voice call either with a live person or

30 on a web platform where prompts are recorded and the subject (potentially with the help of a companion)

31 records their answers. This may be a major advantage of the method over the existing ones that require

32 an in-person interview and use either handcrafted voice features or manual transcriptions. Owing to dealing

1 with text data rather than audio, pre-processing steps such as de-identification, diarization, and sub-test

2 labeling can be conducted efficiently using NLP. In our study, we removed the T (examiner) sentences from

3 the prediction procedure as they are structurally repeated throughout all the interviews, likely containing no

4 useful information to assess the interviewees. Another benefit of removing the T sentences is that a comput-

5 erized version of this framework will only require the participant's responses during a web-based structured

6 NP assessment. Furthermore, the performance of the predictive model improves by taking advantage of

7 the sub-test information. Specifically, in MCI detection, the STS+Dem. method outperforms the RS+Dem.

8 method with a margin of 3.8% in AUC.

¶Comparing the selected sub-tests used in the STS method for each task shows that FAS (verbal fluency)

10 has high impact in identifying the dementia class, while removing it enhances the classification performance

11 in MCI detection. Thus, our approach enables the identification of sub-tests that are more informative for

12 each task. This point highlights that a more structured interview could better capture the language deficits

13 underlying cognitive decline. For instance, given the ranking order of the sub-tests in differentiating MCI

14 from Normal, WAIS-R (general questions) can be more useful for assessment of MCI, whereas FAS would

15 not be as useful in this task, at least from the perspective of generated text our method uses.

16 We observed excellent performance in manual transcriptions, achieving 96% and 94% AUC in Tasks I

17 and II. We further validated the approach on automated transcriptions and obtained encouraging results

18 that can lead to a novel diagnostic tool. There might be several limitations that need to be addressed. The

19 generalization of the proposed methods needs to be validated using different cohorts. In particular, and

20 since clinical speech models tend to be overly optimistic in their reports of accuracy [47], a prospective study

21 with validation of the models in an external data set would further enable us to assess their true accuracy.

22 Unfortunately, this is a major hurdle due to the lack of access to external data, which underscores the need for

23 making data more broadly available. Despite the excellent results in cognitive impairment detection using

24 NP tests administered in English as the spoken language, our approach should be
[1] implemented in other

25 languages to confirm its effectiveness for global use.

[2]

26 NLP models on which our analysis is based, have been found to be very useful in medical research (see,
e.g.,

27 [48, 23] and references therein). However, it is also known that these models may reflect biases
[3] (gender, social,

28 racial, etc.) present in the text corpora used for training the models [49]. At the same time, new
methods

[4]
29 are being introduced that can help mitigate these biases (e.g., [50]). In our approach, the
Framingham NP

30 assessment interview is focusing on topics that are less likely to invoke racial or social biases.
However,

31 modeling biases where the model may exploit superficial features such as the length of input text
should be

[5]
32 taken into account in diagnostic tools. In addition, cultural bias and downstream error have to be
controlled

33 when deploying the proposed pipeline in a region with different culture or dialect.

1 **References**

[6]

Alzheimer's Association, Alzheimer's Disease Facts and Figures, <https://www.alz.org/alzheimers-dementia/facts-figures> (2020).

[7]
P. Scheltens, K. Blennow, M. Breteler, B. de Strooper, G. Frisoni, S. Salloway, W. Van der
Flier, Alzheimer's disease, *Lancet* 388 (2016) 505–517.

R. S. Turner, T. Stubbs, D. A. Davies, B. C. Albeni, Potential new approaches for diagnosis
of Alzheimer's disease and related dementias, *Frontiers in Neurology* 11 (2020) 496.

J. A. Thomas, H. A. Burkhardt, S. Chaudhry, A. D. Ngo, S. Sharma, L. Zhang, R. Au, R. Hosseini Ghomi, Assessing the utility of language and voice biomarkers to predict cognitive impairment in the Framingham heart study cognitive aging cohort data, *Journal of Alzheimer's Disease* 76 (3) (2020) 905–922.

R.-P. Filiou, N. Bier, A. Slegers, B. Houz'e, P. Belchior, S. M. Brambati, Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: a scoping review, *Aphasiology* 34 (6) (2020) 723–755.

B. Downer, D. W. Fardo, F. A. Schmitt, A summary score for the Framingham heart study neuropsychological battery, *Journal of Aging and Health* 27 (7) (2015) 1199–1222.

L. Ashendorf, A. L. Jefferson, R. C. Green, R. A. Stern, Test–retest stability on the WRAT-3 reading subtest in geriatric cognitive evaluations, *Journal of Clinical and Experimental Neuropsychology* 31 (5) (2009) 605–610.

D. Stück, A. Signorini, T. Alhanai, M. Sandoval, C. Lemke, J. Glass, S. Hardy, M. Lavalley, B. Wasserman, T. Ang, et al., Novel digital voice biomarkers of dementia from the Framingham study, *Alzheimer's & Dementia* 14 (7) (2018) P778–P779.

V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, S. F. Cappa, Connected speech in neurodegenerative language disorders: a review, *Frontiers in psychology* 8 (2017) 269.

L. Hernández-Domínguez, S. Ratt'e, G. Sierra-Martínez, A. Roche-Bergua, Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task, *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 10 (2018) 260–268.

L. Liu, S. Zhao, H. Chen, A. Wang, A new machine learning method for identifying Alzheimer's disease, *Simulation Modelling Practice and Theory* 99 (2020) 102023.

[12] M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, C. M. T. González, J. Mekyska, Z. Sm'ekal,

Alzheimer's disease and automatic speech analysis: A review, *Expert Systems with Applications* 150

(2020) 113213.

4 [13] V. Berisha, S. Wang, A. LaCross, J. Liss, Tracking discourse complexity preceding
alzheimer's disease

5diagnosis: a case study comparing the press conferences of presidents Ronald Reagan and George
Herbert

6Walker Bush, *Journal of Alzheimer's Disease* 45 (3) (2015) 959–963.

7 [14] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera,
F. Verhey,

8P. Aalten, P. H. Robert, et al., Automatic speech analysis for the assessment of patients with
predementia

9and Alzheimer's disease, *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1
(1)

10(2015) 112–124.

11 [15] K. C. Fraser, J. A. Meltzer, F. Rudzicz, Linguistic features identify Alzheimer's disease in
narrative

12speech, *Journal of Alzheimer's Disease* 49 (2) (2016) 407–422.

13 [16] M. Asgari, J. Kaye, H. Dodge, Predicting mild cognitive impairment from spontaneous spoken
utterances,

14*Alzheimer's & Dementia: Translational Research & Clinical Interventions* 3 (2) (2017) 219–228.

15 [17] H. Lin, C. Karjadi, T. F. Ang, J. Prajakta, C. McManus, T. W. Alhanai, J. Glass, R. Au,
Identification

16of digital voice biomarkers for cognitive health, *Exploration of Medicine* 1 (2020) 406.

17 [18] J. B. Hessler, M. Stemmler, H. Bickel, Cross-validation of the newly-normed SKT for the
detection of

18 MCI and dementia, *GeroPsych*.

19 [19] L. C. Kourtis, O. B. Regele, J. M. Wright, G. B. Jones, Digital biomarkers for alzheimer's
disease: the

20 mobile/wearable devices opportunity, *NPJ digital medicine* 2 (1) (2019) 1–9.

21 [20] O. G. Iroju, J. O. Olaleke, A systematic review of natural language processing in healthcare,
International

22 *Journal of Information Technology and Computer Science* 7 (8) (2015) 44–50.

23 [21] S. K. Srivastava, S. K. Singh, J. S. Suri, A healthcare text classification system and its
performance

24 evaluation: A source of better intelligence by characterizing healthcare text, in: *Cognitive
Informatics,*

25 *Computer Modelling, and Cognitive Science*, Elsevier, 2020, pp. 319–369.

26 [22] T. B. Murdoch, A. S. Detsky, The inevitable application of big data to health care, *JAMA*
309 (13)

27 (2013) 1351–1352.

1 [23] B. Hao, H. Zhu, I. C. Paschalidis, Enhancing clinical BERT embedding using a biomedical
knowledge

2 base, in: *Proceedings of the 28th International Conference on Computational Linguistics, 2020,*
pp.

3 657–661.

4 [24] C. Andersson, A. D. Johnson, E. J. Benjamin, D. Levy, R. S. Vasan, 70-year legacy of the
Framingham

5 heart study, *Nature Reviews Cardiology* 16 (11) (2019) 687–698.

6 [25] R. Au, R. J. Piers, S. Devine, How technology is reshaping cognitive assessment: Lessons from the

7 Framingham heart study., *Neuropsychology* 31 (8) (2017) 846.

8 [26] A. J. Jak, S. R. Preis, A. S. Beiser, S. Seshadri, P. A. Wolf, M. W. Bondi, R. Au, Neuropsychological

9 criteria for mild cognitive impairment and dementia risk in the Framingham heart study, *Journal of the*

10 International Neuropsychological Society: *JINS* 22 (9) (2016) 937.

11 [27] E. R. McGrath, A. S. Beiser, C. DeCarli, K. L. Plourde, R. S. Vasan, S. M. Greenberg, S. Seshadri,

12 Blood pressure from mid-to late life and risk of incident dementia, *Neurology* 89 (24) (2017) 2447–2454.

13 [28] C. L. Satizabal, A. S. Beiser, V. Chouraki, G. Ch[^]ene, C. Dufouil, S. Seshadri, Incidence of dementia

14 over three decades in the Framingham heart study, *New England Journal of Medicine* 374 (6) (2016)

15 523–532.

16 [29] Cloud speech-to-text API - language support, <https://cloud.google.com/speech-to-text>, accessed:

17 2021-01-30.

18 [30] V. K[^]epuska, G. Bohouta, Comparing speech recognition systems (microsoft api, google api and cmu

19 sphinx), *Int. J. Eng. Res. Appl* 7 (03) (2017) 20–24.

20 [31] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite bert
for self-

21 supervised learning of language representations, arXiv preprint arXiv:1909.11942.

22 [32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional
transformers

23 for language understanding, arXiv preprint arXiv:1810.04805.

24 [33] F. J. Massey Jr, The Kolmogorov-Smirnov test for goodness of fit, *Journal of the American
Statistical*

25 *Association* 46 (253) (1951) 68–78.

26 [34] M. L. McHugh, The chi-square test of independence, *Biochemia medica: Biochemia medica* 23
(2) (2013)

27 143–149.

28 [35] A. Rios, R. Kavuluru, Neural transfer learning for assigning diagnosis codes to EMRs,
Artificial Intelli-

29 *gence in Medicine* 96 (2019) 116–122.

1 [36] M. Lehr, E. Prud'hommeaux, I. Shafran, B. Roark, Fully automated neuropsychological
assessment for

2 detecting mild cognitive impairment, in: *Thirteenth Annual Conference of the International
Speech*

3 *Communication Association*, 2012.

4 [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I.
Polosukhin,

This article is protected by copyright. All rights reserved.

Attention is all you need, arXiv preprint arXiv:1706.03762.

[38] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes,

S. Yuan, C. Tar, et al., Universal sentence encoder, arXiv preprint arXiv:1803.11175.

[39] B. Hao, S. Sotudian, T. Wang, T. Xu, Y. Hu, A. Gaitanidis, K. Breen, G. C. Velmahos, I. C. Paschalidis,

Early prediction of level-of-care requirements in patients with COVID-19, *Elife* 9 (2020) e60519.

[40] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic

(ROC) curve., *Radiology* 143 (1) (1982) 29–36.

[41] M. F. Folstein, S. E. Folstein, P. R. McHugh, “Mini-mental state”: a practical method for grading the

cognitive state of patients for the clinician, *Journal of psychiatric research* 12 (3) (1975) 189–198.

[42] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung,

et al., Multilingual universal sentence encoder for semantic retrieval, arXiv preprint arXiv:1907.04307.

[43] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott,

L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint

arXiv:1911.02116.

This article is protected by copyright. All rights reserved.

19 [44] G. Deshpande, V. S. Viraraghavan, M. Duggirala, S. Patel, Detecting emotional valence
using time-

20domain analysis of speech signals, in: 2019 41st Annual International Conference of the IEEE
Engineering

21in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 3605–3608.

22 [45] B. Köprü, E. Erzin, Multimodal continuous emotion recognition using deep multi-task
learning with

23correlation loss, arXiv preprint arXiv:2011.00876.

24 [46] C. Xue, C. Karjadi, I. C. Paschalidis, R. Au, V. B. Kolachalama, Detection of dementia
on voice

25recordings using deep learning: a framingham heart study, *Alzheimer's Research & Therapy* 13
(1)

26(2021) 146.

27 [47] V. Berisha, C. Krantsevich, P. R. Hahn, S. Hahn, G. Dasarathy, P. Turaga, J. Liss, Digital
medicine

28and the curse of dimensionality, *NPJ digital medicine* 4 (1) (2021) 1–8.

1 [48] Ö. Uzuner, B. R. South, S. Shen, S. L. DuVall, 2010 i2b2/VA challenge on concepts,
assertions, and

2relations in clinical text, *Journal of the American Medical Informatics Association* 18 (5) (2011) 552–
556.

3 [49] S. L. Blodgett, S. Barocas, H. Daum´e III, H. Wallach, Language (technology) is power: A
critical survey

4of "bias" in NLP, arXiv preprint arXiv:2005.14050.

5 [50] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, L.-P. Morency, Towards debiasing sentence

representations, arXiv preprint arXiv:2007.08100.

Table 1: Summary of the variables in the FHS dataset. Diagnosis scores 0, 0.5, 1–1.5, 2–2.5, and 3 are defined as normal cognition, MCI, mild dementia, moderate dementia, and severe dementia, respectively. Education labels 0, 1, 2, and 3 indicate subjects who attended some high school, graduated from high school, attended some college, and graduated from college, respectively. Note that 52 education labels are missing among the Normal group.

Variable	Assessment			<i>p</i> -value		
	Normal (N) (<i>n</i> = 410)	MCI (<i>n</i> = 387)	Dementia (D) (<i>n</i> = 287)	N vs D Task I	N/MCI vs D Task II	N vs MCI Task III
Diagnosis				-	-	-
0	410 (100%)	0 (0%)	0 (0%)			
0.5	0 (0%)	387 (100%)	0 (0%)			
1–1.5	0 (0%)	0 (0%)	180 (63%)			
2–2.5	0 (0%)	0 (0%)	96 (33%)			
3	0 (0%)	0 (0%)	11 (4%)			
Age	77.2 ± 9.7	81.6 ± 8.0	85.1 ± 7.5	<0.001	<0.001	<0.001
Gender				1e-5	8e-5	0.053
Female	204 (50%)	220 (57%)	192 (68%)			
Male	206 (50%)	176 (43%)	95 (32%)			
Education				0.002	0.008	0.289
0	33 (8%)	40 (10%)	37 (13%)			
1	116 (28%)	123 (32%)	108 (38%)			
2	128 (31%)	121 (31%)	71 (25%)			
3	128 (31%)	101 (26%)	68 (24%)			
ApoE				7e-5	0.030	4e-6
22	0 (0%)	2 (0.5%)	5 (2%)			
23	60 (15%)	40 (10%)	25 (9%)			
24	8 (2%)	9 (2%)	5 (2%)			
33	276 (67%)	219 (57%)	170 (59%)			

34	51 (12%)	98 (25%)	64 (22%)
44	4 (1%)	12 (3%)	8 (3%)

Table 2: Results on the test set (mean \pm std over the ten runs). An ensemble model of the sub-test sampling method, random sampling method and demographic information (STS + RS + Dem.) achieves the best performance. 'ApoE' refers to variables indicating the presence of ApoE genes, 'edu.' the use of educational level, and 'MMSE' to Mini-Mental State Examination.

Methods	Tasks	AUC	%Acc	%Sensitivity	%Specificity	%
STS	I	90.2 \pm 4.5	86.6 \pm 3.8	86.2 \pm 4.1	86.9 \pm 7.7	
	II	85.3 \pm 4.3	80.3 \pm 3.9	76.9 \pm 6.2	83.8 \pm 7.1	
	III	71.7 \pm 4.7	69.5 \pm 3.6	65.4 \pm 9.1	73.6 \pm 9.1	
STS+Dem.	I	91.2 \pm 4.1	86.2 \pm 4.1	84.5 \pm 4.7	87.9 \pm 5.4	
	II	87.1 \pm 4.2	81.7 \pm 3.3	78.3 \pm 6.7	85.2 \pm 8.7	
	III	72.8 \pm 4.3	69.4 \pm 3.1	66.2 \pm 12.9	72.6 \pm 8.8	
RS	I	88.5 \pm 3.7	84.0 \pm 2.4	84.1 \pm 6.2	83.8 \pm 5.8	
	II	83.7 \pm 2.4	80.3 \pm 3.0	79.7 \pm 11.2	81.0 \pm 7.1	
	III	68.8 \pm 6.4	66.8 \pm 5.9	72.6 \pm 13.4	61.0 \pm 8.5	
RS+Dem.	I	89.6 \pm 3.6	84.1 \pm 3.3	85.2 \pm 5.6	83.1 \pm 5.4	
	II	84.5 \pm 2.4	80.2 \pm 3.5	79.0 \pm 10.2	81.4 \pm 6.8	
	III	69.0 \pm 6.7	67.1 \pm 5.6	72.6 \pm 11.1	61.5 \pm 7.3	
Dem.	I	74.8 \pm 4.6	72.4 \pm 4.4	75.5 \pm 9.2	69.3 \pm 7.3	
	II	71.1 \pm 4.8	69.0 \pm 4.3	68.3 \pm 8.1	69.7 \pm 7.8	
	III	62.8 \pm 7.9	62.8 \pm 5.6	63.8 \pm 8.2	61.8 \pm 8.7	
Dem.+ApoE	I	77.9 \pm 2.9	74.0 \pm 3.4	79.7 \pm 10.0	68.3 \pm 10.0	
	II	72.3 \pm 5.8	70.3 \pm 5.7	75.2 \pm 8.1	65.6 \pm 8.3	
	III	67.1 \pm 6.4	64.6 \pm 4.7	67.7 \pm 7.2	61.5 \pm 9.8	
Dem.+ApoE + edu.	I	78.1 \pm 2.7	75.5 \pm 2.9	77.9 \pm 10.8	73.1 \pm 8.7	
	II	70.5 \pm 6.3	69.1 \pm 4.3	73.8 \pm 5.4	64.5 \pm 6.9	
	III	67.2 \pm 6.9	63.8 \pm 5.8	66.7 \pm 10.3	61.0 \pm 12.7	
STS+RS+Dem.	I	92.6 \pm 3.3	87.1 \pm 4.0	85.5 \pm 6.1	88.6 \pm 4.6	
	II	88.0 \pm 2.2	83.1 \pm 3.0	83.1 \pm 3.9	83.1 \pm 3.9	
	III	74.1 \pm 4.4	69.5 \pm 4.3	70.3 \pm 9.0	68.7 \pm 6.4	
Full STS+Dem.	I	88.5 \pm 3.7	82.8 \pm 3.6	78.6 \pm 9.9	86.9 \pm 6.9	
	II	83.8 \pm 5.1	81.6 \pm 3.9	82.8 \pm 6.2	80.3 \pm 9.3	
	III	67.8 \pm 3.0	66.2 \pm 3.2	62.1 \pm 10.0	70.3 \pm 11.0	

MMSE	I	82.7 ± 3.9	77.1 ± 2.4	72.4 ± 6.7	81.7 ± 5.8
	II	80.5 ± 4.5	75.2 ± 3.9	73.1 ± 9.2	77.2 ± 9.9
	III	63.5 ± 7.1	62.6 ± 4.5	69.7 ± 8.9	55.4 ± 9.1
STS+Dem.+ApoE	III	74.4 ± 4.4	71.2 ± 3.5	68.2 ± 12.0	73.8 ± 6.5

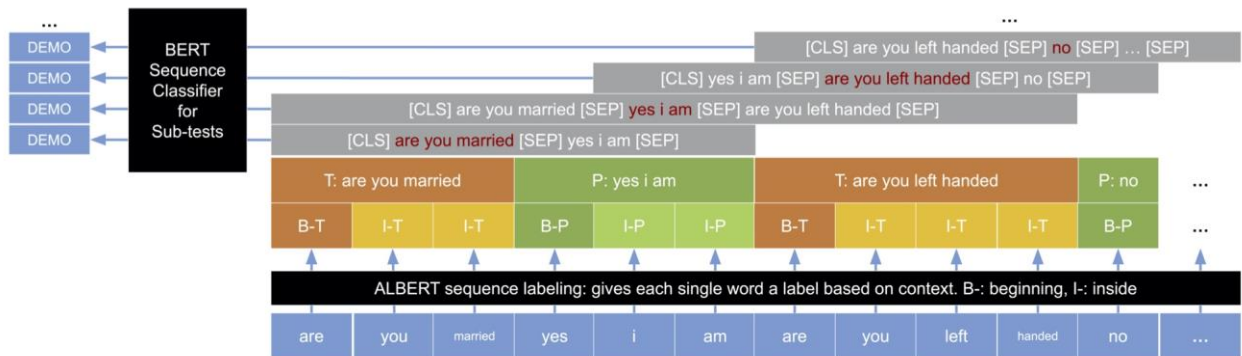


Figure 1: Speaker diarization and NP sub-test labeling using NLP models.

Figure 1: Speaker diarization and NP sub-test labeling using NLP models.

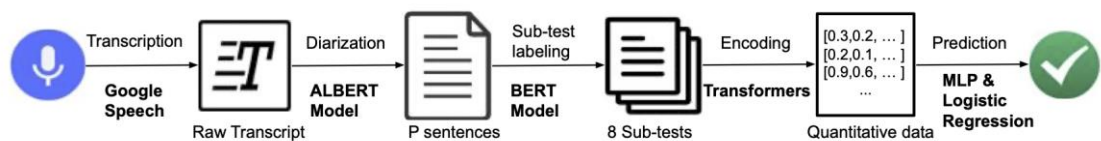


Figure 2: The schematic of the proposed method. We summarized the goal of each step and its corresponding method above and under each arrow, respectively.

Figure 2: The schematic of the proposed method. We summarized the goal of each step and its corresponding method above and under each arrow, respectively.

- (a) The distribution of AUC for different methods in task I.(b) The distribution of AUC for different methods in task II.
- (c) The distribution of AUC for different methods in task III.(d) The average AUC of the STS+RS+Dem. method for each task.

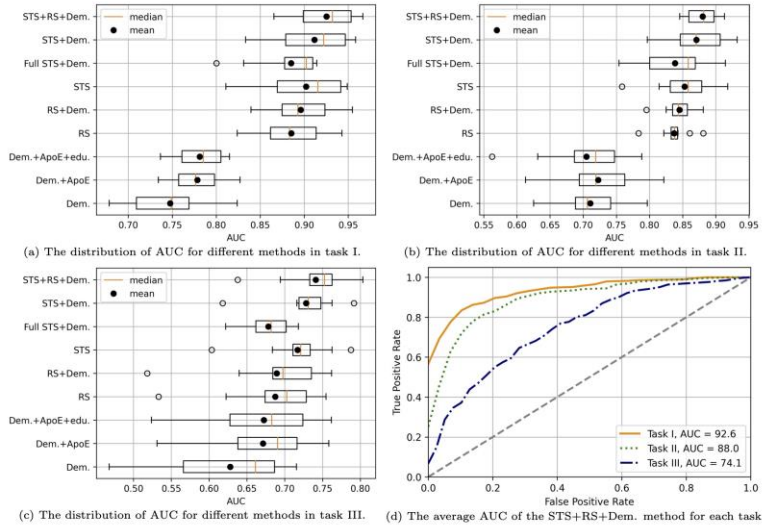


Figure 3: The performance of different methods over 10 splits for each task.

Figure 3: The performance of different methods over 10 splits for each task.

(a) Task I.(b) Task II.(c) Task III.

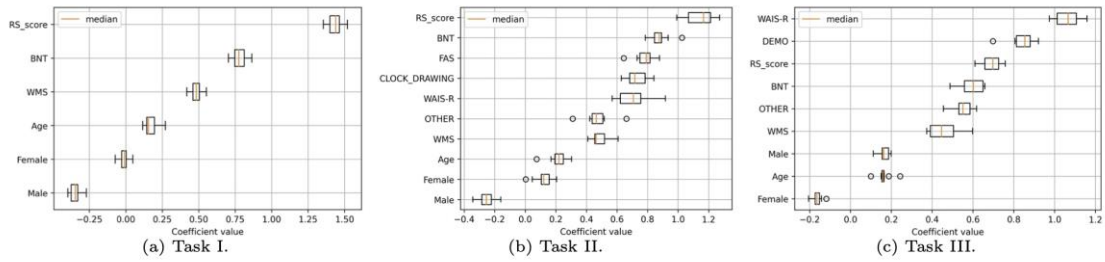


Figure 4: Logistic regression coefficients using the STS+RS+Dem. method, indicating the relative predictive importance of the features in each task.

Figure 4: Logistic regression coefficients using the STS+RS+Dem. method, indicating the relative predictive importance of the features in each task.