



# An Exploration of Interpretable Deep Learning Models for the Assessment of Mild Cognitive Impairment

Emma C. L. Leschly<sup>1,3</sup>, Oliver Roesler<sup>1</sup>, Michael Neumann<sup>1</sup>, Jackson Liscombe<sup>1</sup>, Abhishek Hosamath<sup>1</sup>, Lakshmi Arbatti<sup>1</sup>, Line H. Clemmensen<sup>5</sup>, Melanie Ganz<sup>3,4</sup>, Vikram Ramanarayanan<sup>1,2</sup>

<sup>1</sup>Modality.AI, Inc., San Francisco, CA, USA. <sup>2</sup>University of California, San Francisco, San Francisco, CA, USA. <sup>3</sup>Department of Computer Science, University of Copenhagen, Copenhagen, Denmark. <sup>4</sup>Neurobiology Research Unit, Copenhagen University Hospital, Copenhagen, Denmark. <sup>5</sup>Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark

v@modality.ai

## Abstract

Early diagnosis and intervention are crucial for mild cognitive impairment (MCI), as MCI often progresses to more severe neurodegenerative conditions. In this study, we explore utilizing deep learning for MCI detection without losing the interpretability provided by feature-based approaches. We used a dataset consisting of 90 MCI patients and 91 controls collected via a remote assessment platform and analyzed the participants' spontaneous speech responses to the Patient Report of Problems (PROP) which asks patients to report their most bothersome general health problems. The proposed deep neural network, which features a bottleneck layer including 13 interpretable symptom domains, achieved an AUC of 0.62, thereby outperforming a set of feature-based classifiers while ensuring interpretability due to the bottleneck layer. We further illustrated the model's interpretability by examining how the predicted PROP domains influence final predictions using Shapley values.

**Index Terms:** multimodal dialog system, remote patient monitoring, mild cognitive impairment, interpretability

## 1. Introduction

Mild Cognitive Impairment (MCI) is a neurodegenerative condition characterized by a decline in cognitive functions such as memory, language, and attention that is greater than what is expected based on the individual's age and educational background but not severe enough to significantly impact daily functioning nor meet the criteria for dementia [1]. It is estimated to affect approximately 12% - 18% of people above age 60 and about 10% - 15% of individuals with MCI go on to develop dementia each year, a risk that increases as MCI progresses [2]. Detecting MCI in the early or preclinical stage is crucial in order to start intervention and treatment, potentially slowing down progression [3]. However, because of the heterogeneity of the condition and the subtlety of the cognitive decline, MCI can be difficult to detect, even for experts [4].

Several studies have demonstrated the utility of digital biomarkers that can be measured non-invasively like speaking rate, pitch, or prosody for the assessment of neurological conditions like MCI [5, 6]. For instance, Themistocleous et al. [7] achieved accuracies up to 0.83 using a set of Multi-Layer Perceptrons on speech samples from a reading task, while Fraser et al. [8] combined speech, language, and eye-tracking features using late fusion and obtained an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.88. Bertini et al. [9] utilized autoencoders to extract features from spontaneous speech, achieving F1-scores of 0.85 and 0.91 using

data augmentation, while Vincze et al. [10] classified MCI and mild AD using linguistic features, reaching accuracies of 0.71 and 0.80. Amini et al. [11] demonstrated the effectiveness of transformer-based sentence encoders in predicting progression of MCI to AD, outperforming traditional neuropsychological tests. Finally, Ortiz-Perez et al. [12] developed a multimodal architecture that integrated BERT embeddings and speech features, achieving an Unweighted Average Recall (UAR) of 0.75 during cross-validation; however, performance dropped to 0.56 on an independent test set released later.

Although the results in the presented studies are overall promising, there are several limitations. First, most of the used datasets were relatively small raising the question whether the results are generalizable [13]. Secondly, many of the studies only analyzed a single modality, e.g. text or speech, instead of combining information from multiple modalities. Finally, most of the used datasets were collected in-lab or in-clinic under controlled environments so that the developed models might not work well for data collected in the wild. To address above limitations, we recently collected 362 remote assessments from 90 people with MCI and 91 healthy controls, extracted a number of facial, speech, text, and cognitive features from the collected audio-video data, and provided a set of 13 features from different modalities as input to a support vector machine (SVM) classifier achieving an AUC of 0.75 [14, 15]. While deep learning models have outperformed more traditional feature-based methods on a wide range of tasks, we used the latter in our previous study because the former are usually not interpretable which is critically important for clinical applications.

However, recently Xu et al. [16] proposed a deep neural network (DNN) with a clinically-interpretable bottleneck layer for assessment of dysarthric speech. Their model utilized mel-spectrograms as input which were first processed through two convolutional layers before being passed through an interpretable bottleneck layer mapping the intermediate representation to four acoustic features characterizing dysarthria. The model was trained to jointly learn these features and the final classification label according to a weight parameter  $w$ , where  $w = 1$  and  $w = 0$  put all weight on learning the final label and the features in the bottleneck layer, respectively. Their model improved upon a model without a bottleneck layer and a model that included the bottleneck layer but was only trained to focus on the classification task. Inspired by the work by Xu et al [16], this paper aims to answer the following research questions:

1. How do DL models perform relative to traditional feature-based models despite the usually relatively small datasets available for different neurological conditions?
2. Can the integration of a bottleneck layer provide a similar

level of interpretability as traditional feature-based methods?

To this end, we developed a DNN with a clinically-interpretable bottleneck layer consisting of 13 symptom domains for MCI detection, trained the model to predict the symptom domains and final label using a joint multitask training strategy, and compared its performance to a set of traditional feature-based models. Additionally, we compared the effectiveness of interpretable features with deep learning embeddings as input to the DNN and examined how the predicted symptom domains influence the final predictions using Shapley values [17].

## 2. Data

Our dataset includes a total of 181 participants, 90 (9 female, 81 male) MCI patients and 91 (9 female, 82 male) controls, recruited via the U.S. Department of Veterans Affairs<sup>1</sup> between November 2023 and January 2024. The two cohorts were age-matched with the mean age being 71.08 (SD = 9.1) years for MCI patients and 71.3 (SD = 8.59) years for controls. 90% of participants were male and only 10% were female reflecting the fact that only about 10% of US veterans are female<sup>2</sup>. In the control cohort, 92% identified as white, 5% as black, and 3% as others, while for MCI patients 80% identified as white, 16% as black, and 4% as others. The education levels between cohorts were similar with 44% of the MCI patients and 35% of the controls having an advanced degree, 47% of the patients and 60% of the controls having an undergraduate degree, and 9% of the patients 5% of the controls having a high school degree or General Educational Diploma. The study was approved by the Institutional Review Board of the University of California, San Francisco.

Each participant completed two assessments one week apart administered through a remote assessment platform [18, 19]. In order to qualify, participants had to comply with a set of inclusion and exclusion criteria. They had to be at least 55 years old, able to consent and e-sign, have a valid phone number and email, able to read and speak in English, and have access to a smartphone, tablet, or PC with internet connection and webcam. At the same time, they could not be diagnosed with dementia, had cognitive impairment due to cerebrovascular disease, head trauma, chronic or active abuse of alcohol, opioid, or methamphetamines, Parkinson’s disease, schizophrenia, bipolar disease, or major depressive disorder, or used benzodiazepines, non-BZD receptor modulator sleeping medications, drugs for the treatment of Parkinson’s disease such as levodopa, or antipsychotics. MCI patients had to meet the criteria for at least two MCI diagnoses according to the ICD-10.

Assessments included 23 structured tasks designed to elicit certain speech, facial, and cognitive behaviors. In this study, we chose to focus on a specific spontaneous speech task, namely the Patient Report of Problems (PROP), which comprises of several open-ended questions that asks individuals to report and rank up to five bothersome problems and their impacts on daily functioning [20]. We selected this task because the participants’ responses have direct clinical utility and can therefore be used to obtain ground truth values for the bottleneck layer (see Section 3). Additionally, the PROP responses are ideal for extracting both BERT and wav2vec 2.0 embeddings as responses are natural, individual, and of sufficient duration.

<sup>1</sup><https://www.usa.gov/agencies/u-s-department-of-veterans-affairs>

<sup>2</sup>[https://www.va.gov/vetdata/veteran\\_population.asp](https://www.va.gov/vetdata/veteran_population.asp)

Table 1: Overview of interpretable speech and text features.

Domain	Feature	
Speech	Energy	shimmer, intensity (dB), signal-to-noise ratio (dB)
	Timing	speaking and articulation duration (sec.), speaking and articulation rate (WPM) percentage pause time, canonical timing agreement
	Voice quality	Cepstral peak prominence, CPP (dB) Harmonics-to-noise ratio, HNR (dB)
	Frequency	mean, max., min fundamental frequency F0 (Hz) first three formants F1, F2, F3 (Hz) slope of 2nd formant (Hz/sec.), jitter
Text	Lexical	word count, percentage of content words, noun rate, verb rate, pronoun rate, noun-to-verb ratio, noun-to-pronoun ratio, closed class word ratio
	Semantic	Idea density

## 3. Methods

**Input Features** For each modality, we explored two types of input features for both the baseline and proposed models: interpretable features (Table 1) and contextualized deep learning embeddings. The interpretable text features were obtained by first transcribing the speech recordings using Amazon Transcribe<sup>3</sup> and then computing the features using SpaCy<sup>4</sup>. Speech features were extracted from the original speech recordings using Praat [21] and Kaldi [22]. For text embeddings, we used BERT embeddings [23] using a pre-trained BERT-base-uncased tokenizer and model from Transformers 4.41.2 [24]. The process involved tokenizing the transcriptions, adding special [CLS] and [SEP] tokens, and passing the tokenized input through the model. We then extracted the final hidden state associated with the [CLS] token, resulting in a 768-dimensional vector representing the entire text sequence. For speech embeddings, we used wav2vec 2.0 embeddings<sup>5</sup> [25] using a pre-trained feature extractor, processor, and model from Transformers 4.41.2 [24]. We preprocessed the speech by combining recordings from different parts of the PROP into a single file, standardizing the length to 60 seconds through zero-padding or trimming, and downsampling from 48 kHz to 16 kHz using librosa 0.10.2<sup>6</sup>. The preprocessed speech was then passed through a feature extractor to generate input representations for the pre-trained model. Unlike the text embeddings, where we used the [CLS] token’s final hidden state, we represented the entire speech sample by calculating the mean of the final hidden states across the time dimension, following standard speech processing practices [26].

**Baseline Models** We first performed 10 random splits of 5-fold cross-validation on a set of baselines applying machine learning classifiers available in scikit-learn 1.5.0 [27], including Support Vector Machine (SVM), Logistic Regression (LR), Multi-Layer Perceptron (MLP), and Random Forest (RF), across three modalities: text, speech, and multimodal.

**Proposed Model** We developed a DNN (Figure 1) inspired by the idea of concept bottleneck models [28] and the integration of clinically-interpretable features in the bottleneck layer [16]. The first part consists of encoders that generate embeddings using pre-trained BERT [23] and wav2vec 2.0 [25] models. For the multimodal experiment, the BERT and wav2vec 2.0 embeddings are then concatenated and passed through fully connected layers. The bottleneck layer that maps the intermediate representations to the concepts is followed by a fully con-

<sup>3</sup><https://aws.amazon.com/transcribe/>

<sup>4</sup><https://spacy.io/>

<sup>5</sup><https://huggingface.co/facebook/wav2vec2-base-960h>

<sup>6</sup><https://zenodo.org/records/4923181>

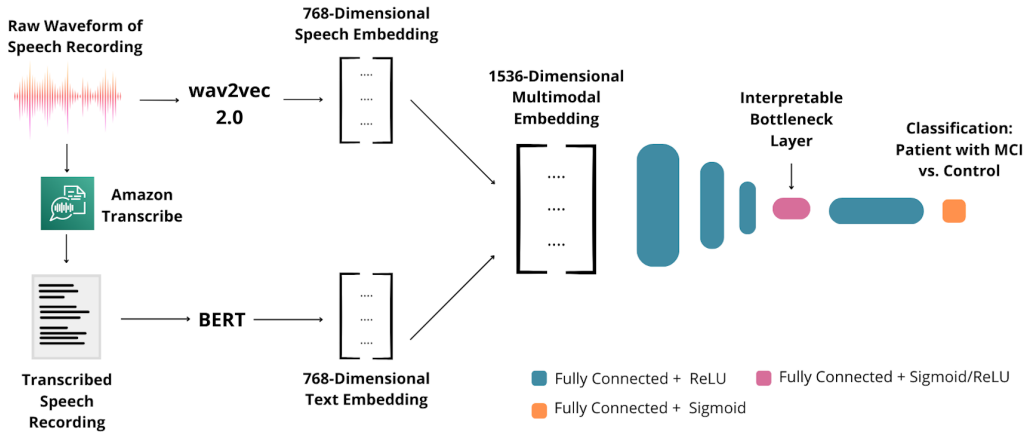


Figure 1: The multimodal version of the DNN architecture. First, the speech recording is transcribed using Amazon Transcribe. Then, contextualized embeddings are extracted from the transcription using BERT and the speech recording using wav2vec 2.0. The two embeddings are concatenated to form a 1536-dimensional vector which is passed through three fully connected layers, a bottleneck layer that maps the intermediate representation to the interpretable features, a fully connected layer, and finally an output layer with a sigmoid activation for binary classification into MCI or control.

ected layer. Finally, the output layer predicts the binary label. The model is trained using a joint, multitask training strategy that employs a custom loss function that combines binary cross-entropy for final classification and prediction of the bottleneck concepts:

$$L_{total} = w \cdot L_{classification} + (1 - w) \cdot L_{bottleneck} \quad (1)$$

where  $w$  is a weighting factor balancing the importance of the binary classification and the predicting the bottleneck features. The ground truths of the concepts in the bottleneck layer are one-hot encodings of 13 clinically validated symptom domains, including Pain, Gait, Other Motor, Psychiatric, Cognition, Autonomic Dysfunction, Sleep, Postural Instability, Rigidity, Fatigue, Bradykinesia, Tremor, and Fluctuations. Each domain covers a number of symptoms which are inferred from the self-reported problems using a neural network with two hidden layers developed by Marras et al. [29] who trained it on 168,260 self-reported problems collected from about 25,000 Parkinson’s Disease patients and achieved an accuracy of 95%.

To investigate the impact of different input modalities, we conducted the following experiments:

1. *Text*: Using 768-dimensional BERT embeddings extracted from transcriptions of the PROP responses.
2. *Speech*: Using 768-dimensional wav2vec 2.0 embeddings extracted from the first 60 seconds of the speech recordings.
3. *Multimodal*: Using 1536-dimensional multimodal embeddings (text and speech) obtained through concatenation of BERT and wav2vec 2.0 embeddings.

We performed hyper-parameter optimization using hyperopt 0.2.7 [30] on the following hyper-parameters: learning rate (log-uniformly sampled between  $1e-5$  and  $1e-2$ ), layer dimensions ([256], [512, 256], or [512, 256, 128] with an optional layer of 32, 64, or 128 neurons), dropout rate (uniformly sampled between 0.1 and 0.3), batch normalization (True or False), batch size (16, 32, or 64), optimizer (Adam, SGD with momentum, or RMSprop), and learning rate scheduler (None, ReduceLROnPlateau, or StepLR); each model was optimized separately over 10 epochs and a maximum of 100 evaluations, using AUC on the validation set as the optimization metric.

Following hyperparameter optimization, the models were trained for 50 epochs. Our model training was conducted on

an NVIDIA Tesla T4 GPU with 16 GB of memory. To ensure robust model evaluation and mitigate data leakage, we used a stratified group 5-fold cross-validation approach. During each iteration, we collected predictions from each fold and computed metrics only after all folds had been evaluated. This allowed us to calculate scores on the entire dataset, rather than averaging scores across folds. We repeated the cross-validation 10 times using different random seeds to measure the variance of the model’s performance across different splits to obtain a more robust evaluation following guidelines from Varoquaux et al. [31].

## 4. Results

**Classification** Table 2 shows the classification results across different modalities and values of  $w$ . The multimodal DNN achieved the best performance with a peak AUC of 0.620 at  $w = 0.9$  and a peak accuracy, precision, and recall of 0.588, 0.591, and 0.588, respectively, at  $w = 0.8$ . The performances of the multimodal DNN at  $w = 0.4, 0.5, 0.6, 0.7$  and  $0.8$  are not statistically significantly different from that of the configuration at  $w = 0.9$ , based on a Bonferroni-corrected Dunn’s test following a non-parametric Kruskal-Wallis test that showed statistically significant differences among the groups ( $p < 0.001$ ). Notably, the multimodal DNN outperformed both the text ( $p = 0.004$ ) and speech ( $p = 0.03$ ) DNNs. The text DNN yielded the highest performance at  $w = 0.5$  with an accuracy of 0.554 and AUC of 0.546, whereas the speech DNN showed the highest performance at  $w = 0.9$  with an accuracy of 0.555 and AUC of 0.580. The precision and recall values generally followed similar trends to accuracy and AUC across different  $w$  values for all DNNs.

The proposed model showed a statistically significant improvement over the best baseline model, logistic regression using multimodal embeddings ( $p = 0.008$ ), while for the baseline models multimodal embeddings performed significantly better than the interpretable speech features ( $p = 0.002$ ) (Table 3). However, the modest effect size underscores the inherent complexity of the task. Generally, the performance of the DNN models increased with  $w$  because more weight is put on learning to distinguish MCI patients from controls. However, the best performance was not achieved at  $w = 1.0$  suggesting that only focusing on the final classification may overlook valuable information encoded in the bottleneck features, which are es-

Table 2: Results of DNN experiments using PROP domains as the bottleneck layer on the binary classification task of MCI/control. The table shows the average Accuracy (Acc.), Precision (Prec.), Recall (Rec.), and Area Under the Curve (AUC) obtained across 10 splits of 5-fold cross-validation for different values of  $w$ . Scores that did not show statistically significant differences to the best AUC, according to a Bonferroni-corrected Dunn’s test following a Kruskal-Wallis test, are marked in bold.

$w$	Text				Speech				Multimodal			
	Acc.	Prec.	Rec.	AUC	Acc.	Prec.	Rec.	AUC	Acc.	Prec.	Rec.	AUC
0.1	0.532	0.512	0.532	0.508	0.505	0.478	0.505	0.457	0.518	0.526	0.518	0.529
0.2	0.5217	0.539	0.517	0.487	0.505	0.461	0.505	0.458	0.518	0.529	0.518	0.539
0.3	0.534	0.544	0.534	0.510	0.514	0.493	0.514	0.466	0.536	0.547	0.536	0.552
0.4	0.533	0.560	0.533	0.511	0.513	0.509	0.513	0.497	0.554	0.568	0.554	<b>0.575</b>
0.5	0.554	0.591	0.554	0.546	0.536	0.534	0.536	0.540	0.559	0.567	0.559	<b>0.591</b>
0.6	0.539	0.536	0.539	0.518	0.525	0.518	0.525	0.513	0.583	0.591	0.583	<b>0.611</b>
0.7	0.541	0.508	0.541	0.515	0.551	0.552	0.551	0.563	0.581	0.585	0.581	<b>0.616</b>
0.8	0.532	0.501	0.532	0.506	0.553	0.553	0.553	0.563	0.588	0.590	0.588	<b>0.619</b>
0.9	0.529	0.485	0.529	0.502	0.555	0.561	0.555	0.580	0.586	0.587	0.586	<b>0.620</b>
1.0	0.516	0.288	0.516	0.445	0.545	0.544	0.545	0.564	0.510	0.358	0.510	0.455

Table 3: AUC values averaged across 10 splits of 5-fold cross-validation for each baseline experiment. The scores that did not show statistically significant differences to the best AUC of 0.58, according to a Bonferroni-corrected Dunn’s test following a Kruskal-Wallis test, are marked in bold.

Experiment	SVM	LR	MLP	RF
Text features	0.51	<b>0.53</b>	0.51	0.51
BERT embeddings	<b>0.57</b>	<b>0.57</b>	<b>0.56</b>	<b>0.56</b>
Speech features	0.49	0.51	0.50	<b>0.56</b>
wav2vec 2.0 embeddings	0.45	<b>0.53</b>	<b>0.54</b>	<b>0.52</b>
Multimodal features	0.51	<b>0.52</b>	<b>0.54</b>	<b>0.55</b>
Multimodal embeddings	<b>0.52</b>	<b>0.58</b>	<b>0.58</b>	<b>0.57</b>

essential to capture the underlying complexities of the data. Shifting some focus away from optimizing the final prediction may also act as a form of regularization to help avoid overfitting.

**Shapley values** Using the DeepExplainer from shap 0.46.0 [32], we calculated the Shapley values for the bottleneck features to understand their impact on the final output of the best model ( $w = 0.8$ ). We selected a representative background dataset of 100 samples from the training data. By averaging the Shapley values across all samples, we obtained a measure of the average contribution of each bottleneck feature to the model’s predictions. We then sorted the features based on their absolute Shapley values to identify the most influential ones on the final prediction (Figure 2). We observed that Pain had the highest Shapley value indicating that this symptom domain had the strongest influence on the model’s predictions. This was followed by Gait and Cognition suggesting these features also played significant roles in distinguishing MCI from controls. All averaged Shapley values were positive and relatively small suggesting that increases in these features generally contributed to a higher likelihood of MCI classification but also that their influence was modest overall.

## 5. Discussion

The results show that combining text and speech in the form of BERT and wav2vec 2.0 embeddings performs statistically significantly better than either of the modalities alone when given as input to the proposed DNN. However, when the embeddings are given as input to a less complex model such as Logistic Regression, we only observe statistically significant differences between speech and multimodal embeddings. Additionally, the proposed DNN model achieved better performance than the baseline models. Finally, the introduction of the bot-

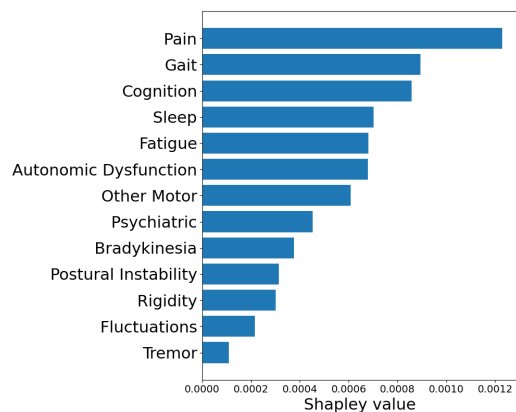


Figure 2: Shapley values for the 13 domains, computed for the multimodal DNN with  $w = 0.8$ , indicating their impact on the final output. Higher values indicate larger influence.

tleneck layer did not only enhance the interpretability of the model by indicating which symptom domains are most relevant to distinguish MCI patients and controls but also improved the classification performance of the DNN model.

While the results show that DL models outperform traditional feature-based models despite the relatively small dataset and can provide a similar level of interpretability when using a bottleneck layer, there are several limitations and directions for future work. First, participants were only recruited from the US veteran population with its specific characteristics and while the number of participants was higher than in many other studies it might still be relatively small to fully leverage the potential of DL models. Second, we only used a single task and two modalities, i.e. text and speech, in this study, although other tasks and features extracted from video have shown the strongest signal in a previous study that used the same dataset. Finally, we did not fine-tune the pre-trained BERT and wav2vec 2.0 models on domain-specific data, which could potentially enhance their ability to capture subtle language and speech changes associated with MCI. In future work, we will explore the use of different concepts in the bottleneck layer and alternative training strategies, such as sequential or independent strategies, to potentially improve upon the joint multitask strategy. Furthermore, we will evaluate our approach on conditions where speech impairments are more pronounced, such as Parkinson’s Disease, and also evaluate its utility in longitudinal studies to track the progression of speech and language changes over time.

## 6. References

- [1] R. C. Petersen, "Mild Cognitive Impairment:," *CONTINUUM: Lifelong Learning in Neurology*, vol. 22, no. 2, Dementia, 2016.
- [2] R. C. Petersen, O. Lopez, M. J. Armstrong, T. S. Getchius, M. Ganguli, D. Gloss, G. S. Gronseth, D. Marson, T. Pringsheim, G. S. Day, M. Sager, J. Stevens, and A. Rae-Grant, "Practice guideline update summary: Mild cognitive impairment: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology," *Neurology*, vol. 90, no. 3, pp. 126–135, Jan. 2018.
- [3] A. Wallin, A. Nordlund, M. Jonsson, K. Lind, Edman, M. Göthlin, J. Stålhammar, M. Eckerström, S. Kern, A. Börjesson-Hanson, M. Carlsson, E. Olsson, H. Zetterberg, K. Blennow, J. Svensson, A. Öhrfelt, M. Bjerke, S. Rolstad, and C. Eckerström, "The Gothenburg MCI study: Design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up," *Journal of Cerebral Blood Flow & Metabolism*, vol. 36, no. 1, Jan. 2016.
- [4] L. Boise, M. B. Neal, and J. Kaye, "Dementia Assessment in Primary Care: Results From a Study in Three Managed Care Systems," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 59, no. 6, Jun. 2004.
- [5] V. Boschi, E. Catricalà, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected Speech in Neurodegenerative Language Disorders: A Review," *Frontiers in Psychology*, vol. 8, 2017.
- [6] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, and J. R. Green, "Speech as a Biomarker: Opportunities, Interpretability, and Challenges," *Perspectives of the ASHA Special Interest Groups*, vol. 7, no. 1, pp. 276–283, Feb. 2022.
- [7] C. Themistocleous, M. Eckerström, and D. Kokkinakis, "Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks," *Frontiers in Neurology*, vol. 9, p. 975, Nov. 2018.
- [8] K. C. Fraser, K. Lundholm Fors, M. Eckerström, F. Öhman, and D. Kokkinakis, "Predicting MCI Status From Multimodal Language Data Using Cascaded Classifiers," *Frontiers in Aging Neuroscience*, vol. 11, p. 205, Aug. 2019.
- [9] F. Bertini, D. Allevi, G. Lutero, D. Montesi, and L. Calzà, "Automatic Speech Classifier for Mild Cognitive Impairment and Early Dementia," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–11, Jan. 2022.
- [10] V. Vincze, M. K. Szabó, I. Hoffmann, L. Tóth, M. Pákási, J. Kálmán, and G. Gosztolya, "Linguistic Parameters of Spontaneous Speech for Identifying Mild Cognitive Impairment and Alzheimer Disease," *Computational Linguistics*, vol. 48, no. 1, pp. 119–153, Apr. 2022.
- [11] S. Amini, B. Hao, J. Yang, C. Karjadi, V. B. Kolachalama, R. Au, and I. C. Paschalidis, "Prediction of alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models," *Alzheimer's & Dementia*, vol. 20, no. 8, 2024.
- [12] D. Ortiz-Perez, J. Garcia-Rodriguez, and D. Tomás, "Cognitive Insights Across Languages: Enhancing Multimodal Interview Analysis," Jun. 2024, arXiv:2406.07542 [cs].
- [13] V. Berisha, C. Krantsevich, G. Stegmann, S. Hahn, and J. Liss, "Are reported accuracies in the clinical speech machine learning literature overoptimistic?" in *Interspeech 2022*, Sep. 2022.
- [14] O. Roesler, J. Liscombe, M. Neumann, H. Kothare, A. Hosamath, L. Arbatti, D. Habberstad, C. Suendermann-Oeft, M. Bartlett, C. Zhang, N. Sukhdev, K. Wilms, A. Badathala, S. Istas, S. Ruhmel, B. Hansen, M. Hannan, D. Henley, A. Wallace, I. Shoulson, D. Suendermann-Oeft, and V. Ramanarayanan, "Towards Scalable Remote Assessment of Mild Cognitive Impairment Via Multimodal Dialog," in *Proceedings of Interspeech*, 2024.
- [15] A. McGarry, O. Roesler, J. Liscombe, M. Neumann, H. Kothare, A. Hosamath, L. Arbatti, A. Badathala, S. Ruhmel, B. J. Hansen et al., "Much more than the malady: The promise of a web-based digital platform incorporating self-report for research and clinical care in mild cognitive impairment," *Mayo Clinic Proceedings: Digital Health*, p. 100224, 2025.
- [16] L. Xu, J. Liss, and V. Berisha, "Dysarthria detection based on a deep learning model with a clinically-interpretable layer," *JASA Express Letters*, vol. 3, no. 1, Jan. 2023.
- [17] L. S. Shapley, *Notes on the N-Person Game &mdash; II: The Value of an N-Person Game*. RAND Corporation, 1951.
- [18] V. Ramanarayanan, D. Pautler, I. Shoulson, L. Arbatti, A. Hosamath, M. Neumann, H. Kothare, O. Roesler, J. Liscombe, A. Cornish, D. Habberstad, V. Richter, D. Fox, D. Suendermann-Oeft, and I. Shoulson, "When Words Speak Just as Loudly as Actions: Virtual Agent Based Remote Health Assessment Integrating What Patients Say with What They Do," 2023, pp. 678–679.
- [19] V. Ramanarayanan, "Multimodal technologies for remote assessment of neurological and mental health," *Journal of Speech, Language, and Hearing Research*, pp. 1–8, 2024.
- [20] L. Arbatti, A. Hosamath, V. Ramanarayanan, and I. Shoulson, "What Do Patients Say About Their Disease Symptoms? Deep Multilabel Text Classification With Human-in-the-Loop Curation for Automatic Labeling of Patient Self Reports of Problems," May 2023, arXiv:2305.04905 [cs, eess].
- [21] P. Boersma, "Praat, a system for doing phonetics by computer," *Glott. Int.*, vol. 5, no. 9, pp. 341–345, 2001.
- [22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.
- [24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020.
- [25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [26] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-Supervised Speech Representation Learning: A Review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, Oct. 2022.
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
- [28] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept Bottleneck Models," Dec. 2020, arXiv:2007.04612 [cs, stat].
- [29] C. Marras, L. Arbatti, A. Hosamath, A. Amara, K. E. Anderson, L. M. Chahine, S. Eberly, D. Kinel, S. Mantri, S. Mathur, D. Oakes, J. L. Purks, D. G. Standaert, C. M. Tanner, D. Weintraub, and I. Shoulson, "What Patients Say: Large-Scale Analyses of Replies to the Parkinson's Disease Patient Report of Problems (PD-PROP)," *Journal of Parkinson's Disease*, vol. 13, no. 5, pp. 757–767, Jul. 2023.
- [30] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, no. 1, Jun 2013.
- [31] G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion, "Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines," *NeuroImage*, vol. 145, pp. 166–179, Jan. 2017.
- [32] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.