2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 979-8-3315-4426-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/ASRU65441.2025.11434604

Recognizing Dementia from Neuropsychological Tests with State Space Models

Liming Wang* Saurabhchand Bhati* Cody Karjadi† Rhoda Au† James Glass*
*Massachusetts Institute of Technology †Boston University
limingw@csail.mit.edu

***Abstract*— Early detection of dementia is critical for timely medical intervention and improved patient outcomes. Neuropsychological tests are widely used for cognitive assessment but have traditionally relied on manual scoring. Automatic dementia classification (ADC) systems aim to infer cognitive decline directly from speech recordings of such tests. We propose Demenba, a novel ADC framework based on state space models, which scale linearly in memory and computation with sequence length. Trained on over 1,000 hours of cognitive assessments administered to Framingham Heart Study participants, some of whom were diagnosed with dementia through adjudicated review, our method outperforms prior approaches in fine-grained dementia classification by 21%, while using fewer parameters. We further analyze its scaling behavior and demonstrate that our model gains additional improvement when fused with large language models, paving the way for more transparent and scalable dementia assessment tools.**[1][2]

***Index Terms*—Dementia classification, speech biomarkers, pathological speech processing, state-space model**

I. Introduction

Dementia, including Alzheimer's disease, is a progressive neurodegenerative condition that severely impairs memory, reasoning, communication, and other cognitive functions. As the global population ages, early diagnosis of dementia has become increasingly important for initiating timely treatment, slowing disease progression, and improving patients' quality of life and life expectancy [1], [2].

A common clinical approach for assessing cognitive impairment is the neuropsychological test, such as the widely used Mini-Mental State Examination (MMSE) [3]. In these tests, trained clinicians guide patients through a series of structured tasks evaluating memory, attention, perception, verbal fluency and reasoning. While effective, this process is labor-intensive, prone to subjective biases and often inconsistent, especially in distinguishing subtle or early-stage cognitive decline [4]. Furthermore, manual scoring may miss nuanced acoustic or linguistic indicators that signal cognitive deterioration. To alleviate these issues, automatic dementia classification (ADC) systems have been developed to infer cognitive status directly from neuropsychological tests [4]–[6]. By analyzing speech recordings, these systems aim to detect subtle linguistic and paralinguistic cues (e.g., hesitations, disfluencies, semantic anomalies) indicative of dementia. In addition to reducing clinician burden, such systems offer the potential to standardize assessment and mitigate biases, including those introduced by leading or suggestive questions [6]–[8].

However, existing ADC systems are fundamentally limited in their ability to process long neuropsychological test recordings — typically around one hour in duration. Most current models, particularly those based on transformer architectures [9], struggle to handle more than 30 seconds of audio at a time due to quadratic growth in memory and computation with input length [10]. This constraint often forces segment-level inference using forced alignment or manual heuristics [6], [11], [12], leading to context fragmentation and a drop in fine-grained classification performance [13]. Alternatively, transcription-based pipelines using an automatic speech recognizer (ASR) suffer from loss of acoustic information as well as error propagation, especially in noisy, spontaneous and multi-speaker conversational settings.

To address these challenges, we propose leveraging state-space models (SSMs) [10], [14], [15], a family of architectures designed for efficient long-sequence modeling. Unlike transformers, SSMs scale linearly in both space and time, making them ideal for modeling full-length interviews without segmentation. Furthermore, the dementia information in the neuropsychological tests can be subtle and sporadic, with many conversational turns offering little diagnostic value [7]. SSMs' natural capacity for temporal compression allows them to distill salient patterns with minimal information loss, making them particularly well-suited for the ADC task. In this paper, we make three main contributions:

1) We present *Demenba*, a memory- and compute-efficient architecture trained on over 1,000 hours of neuropsychological tests with balanced representation across dementia stages;
2) Our method achieves superior ADC accuracy compared to the state-of-the-art model [6] by up to 21% relative AUC, particularly in fine-grained classification settings (e.g., mild cognitive impairment vs. dementia). It also requires significantly fewer trainable parameters. The performance of our method further improves after fusing with text-based large language models (LLMs);
3) We conduct extensive ablation studies to assess the contribution of different speech segments. Our scaling

[1]Code: https://github.com/lwang114/Demenba

[2]This work was supported by the Framingham Heart Study's National Heart, Lung, and Blood Institute contract N01-HC-25195; National Institutes of Health grants U19-AG068753, R01- AG016495, R01-AG008122, R01-AG033040. The authors would also like to thank the staff and participants of the Framingham Heart Study.

experiments further suggest that Demenba maintains robust performance as both data and model size change.

II. Related work

Early work on ADC tasks such as Alzheimer detection (AD) used classical machine learning algorithms with hand-crafted speech and linguistic features [5], [16]. More recent systems leverage deep learning architectures such as convolutional [6] and recurrent [17], [18] neural networks and neural embeddings from pretrained speech representation models such as wav2vec 2.0 [11], [19] and Whisper [12], [20] as well as text language models [12], [21]. Despite progress in algorithmic design, existing work still focuses on sentence-level speech segments and small datasets such as Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) [4], [16] and Fragmingham Heart Study (FHS) 92-hour subset [5] with less than 30 hours of dementia speech combined due to privacy concerns. Among the works, [6]–[8] included examiner speech in their study and surprisingly found that AD is possible with examiner speech only, indicating examiner bias.

State-space models (SSMs) are proposed as more memory- and compute-efficient alternatives to transformers [9], especially when processing long sequences [10], [14], [15]. While earlier SSMs are auto-regressive with linear recurrent layers [10], later variants improve their expressivity by incorporating data-dependent selective scan block [15], [22]. To further increase their modeling capacity, particularly of two-dimensional data such as vision, bidirectional connections from left-to-right [23] and multidirectional connections from left-to-right and top-to-bottom [22] have been proposed. SSMs have been successfully applied to the audio signal in tasks such as audio event classification [13], [24], [25] and self-supervised audio representation learning [26]. Knowledge distillation [13] from transformer teachers can help boost SSMs' performance and even outperform the transformer teachers. SSMs have also been used for efficient automatic speech recognition [27], [28], separation, and synthesis [28].

III. Method

The overall architecture of our SSM-based ADC system is illustrated in Fig. 1. The complete system consists of four main components: a speech segmenter, an automatic speech recognizer (ASR), an audio-based dementia classifier, and a text-based dementia classifier. First, the speech segmenter divides each hour-long recording into shorter chunks using either a simple voice activity detector (VAD) or a more fine-grained speaker diarization model (SD). To preserve longer context while respecting a fixed memory budget, adjacent chunks are merged into segments up to a predefined maximum length (e.g., six minutes). Each merged segment is represented by its mel-filterbank, which serve as input to the audio classifier. To handle long segments, we employ Mamba [15] SSM backbone, retaining the exactly the same model parameters as in the setup of [13]. Concretely, the backbone consists of four groups of Mamba layers, each followed by a downsampling factor of two; the final output

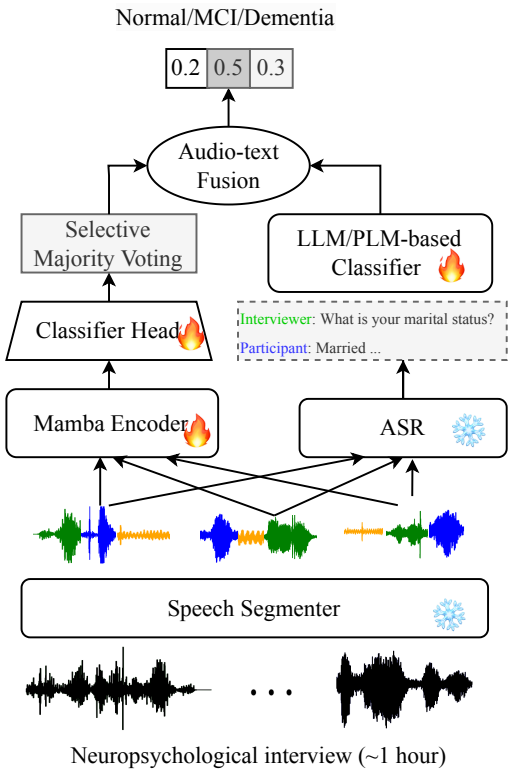


Fig. 1: **Overall architecture of the proposed ADC model**. The model consists of a frozen speech segmenter that divides the hour-long recording into shorter segments, a trainable SSM-based audio classifier and a trainable text-based text classifier. The predictions from the two classifiers are then combined via late fusion.

is mean-pooled to produce class logits. Training is performed with a cross entropy loss. During inference, we observe that treating every segment equally (i.e., standard majority voting) allows noisy or uninformative segments to dilute the overall decision. To mitigate this, we adopt a *selective* majority voting scheme. After computing segment-level probabilities $p^1_{\text{audio}}, \cdots, p^N_{\text{audio}}$ for all N segments, we choose the top k segments by highest peak probability (where k is tuned on a held-out validation set). A soft majority vote over these k segments using the estimated class probabilities produce the recording-level score p_{audio}. In preliminary experiments, this selective voting consistently outperformed naïve majority voting over all segments by filtering out segments that are neither indicative of dementia nor contain reliable audio cues.

In parallel, the SD output is fed to the ASR to obtain speaker-tagged transcripts: for each utterance, we prepend the speaker label (e.g., "Participant" or "Interviewer") to the transcribed text, yielding a complete transcript for the recording. The text classifier then consumes this concatenated transcript to produce a text-based probability p_{text}. We explore both zero-shot classification with large language models (LLMs) and finetuning approaches using either LLMs or pretrained LMs (PLMs) such as BERT [29]. For zero-shot LLM inference, we

construct a structural prompt that includes:

- *Task description*: "You are a helpful assistant that classifies if a participant in an interview has dementia";
- *Interview transcript*: each line formatted as `[speaker role]:[transcript]`, where `[speaker role]` is "Participant" or "Interviewer";
- *Label set*: a list valid labels for the task, e.g., "normal, dementia" for 2-class classification.
- *Instruction*: simply "Answer:".

We then extract the logit corresponding to the first token of the model's response and map it to a probability score over the two labels. Because most LLMs have a strict context-length limit, we either (a) feed the entire transcript into the model at once – if it fits within the context window – or (b) apply the same selective majority voting scheme as in the audio branch. For fine-tuning experiments (both with LLMs and PLMs), we only use selective majority voting to reduce GPU memory usage: we fine-tune on individual segments, select the top k segment scores at inference time, and then combine them. Lastly, we fuse the audio and text scores via

$$p_{\text{audio+text}} = (1-\lambda)p_{\text{audio}} + \lambda p_{\text{text}}, \tag{1}$$

where the *fusion weight* $\lambda \in [0, 1]$. This late fusion allows the system to leverage complementary evidence from both modalities while controlling for over-reliance on either branch.

IV. Experiment

We evaluate the models on tasks including dementia detection and the more fine-grained 3-class dementia classification on the Framingham Heart Study (FHS) dataset [7]. The details are described in the next sections.

TABLE I: Statistics of the FHS dataset. N/M/D stands for the number of recordings for normal/MCI/dementia participants.

split	# participants	# interviewers	N/M/D	# hours
train (2-class)	378	61	400/234/173	943
train (3-class)	586	74	407/399/399	1447
test (2-class)	10	11	10/5/5	25
test (3-class)	18	11	10/10/10	36
eval	92	20	68/10/14	77

A. Dataset and training settings

We conduct all experiments on the Framingham Heart Study (FHS) dataset [7], which consists of approximately 11,000 hour-long neuropsychological test recordings, 2,058 of which have been reviewed for dementia. After adjudication, each recording is labeled as one of three classes — normal, mild cognitive impairment (MCI) and dementia — of which 936 are normal. To balance the dataset, we randomly select roughly 400 recordings per class, yielding 1,200 recordings in total. We consider two classification settings:

- *3-class* classification using the original labels.
- *2-class* classification by merging MCI and dementia into a single "impaired" class and randomly sampling 400 recordings from the combined 800 recordings.

In both settings, the test set comprises 10 recordings per class, sampled randomly. We ensure that participants do not overlap between the training, test splits. Additional dataset statistics are provided in Table I. To segment each recording, we employ two approaches:

- A coarse-grained segmentation using a VAD[3] to split the recordings into silence and speech segments.
- A fine-grained segmentation using an SD system from the pyannote toolkit [30], [31].

In the latter approach, we feed the entire speech waveform into the diarizer and label the first detected speaker with a segment duration more than 20ms as the interviewer, and treat all the subsequent speakers as the participant. For both methods, we merge consecutive segments — up to a maximum duration of 360 seconds — to form the model input. For text-based classifiers, we apply Whisper-Large v2 ASR [20] to each diarized speech segment. To compare with prior work, we also evaluate our approach on a disjoint 92-recording subset of FHS [5], [6] that contains manual transcripts and speaker diarization labels. On this subset, the ASR yields a character error rate of approximately 60%, reflecting prevalent disfluencies and noisy recording conditions. We also inspected the frequency of dementia-related keywords in our training and test sets, finding 83 occurrences of "Alzheimer" and 72 occurrences of "dementia" in the training set, and only one occurrence of "Alzheimer" and zero occurences "dementia" in the test set. By examining the contexts of these mentions, we confirmed that none directly reveal participants' cognitive status, consistent with standard neuropsychological testing protocols that prohibit interviewers from disclosing such information [7]. We follow prior work [6] and extract 128-bin mel-filterbanks with a 10ms frame shift, a 25ms frame length, and a Hanning window. Our audio classifier is based on VMamba [22], configured exactly as in DASS [13] and initialized with ImageNet [32]-pretrained weights, which outperform an AudioSet [33]-pretrained alternative in preliminary experiments. We refer to the larger model as Demenba-medium, corresponding to DASS-medium [13], and to the smaller model as Demenba-small, sharing its architecture with DASS-small [13]. For comparison, we also implement the previous state-of-the-art EfficientNet-based model [6], using the b6 variant [34], which is the largest variant that fits our GPU memory constraints when processing 360-second segments. All models share the same input features, segment durations, and optimization hyperparameters such as batch size and training epochs. For the text classifier, we experiment with two families of text models: PLMs such as BERT [29], with a three-layer multilayer perceptron (MLP) of 768 hidden units appended on top of BERT's final embedding layer, and LLMs, including Llama [35], Qwen2 [36] and phi-4 [37]. For each LLM, we apply a low-rank adaptor (LoRA) [38] of rank 8 to every layer for finetuning. In the 2-class setting, we compare binary cross entropy (BCE) and (multinomial) cross entropy (CE) losses, finding that CE works best for

[3]https://github.com/wiseman/py-webrtcvad.git

TABLE II: **Dementia classification results for audio-only models**. We compare the proposed methods with the state-of-the-art method based on EfficientNet [6]. only AUCs for the better of BCE and CE losses are shown. All models are trained on the 400 hour/class datasets with mel-filterbank features of 360-second segments as inputs. $(3 \rightarrow 2)$ means results from a 3-class classifier by merging MCI+dementia probabilities.

	# Trainable Param.	Segment Boundary	Loss	AUC (↑)
2-class Classification				
EfficientNet b6	40m	VAD	BCE	0.76
		SD	BCE	0.82
Demenba-small	29m	VAD	CE	0.77
		SD	CE	<u>0.85</u>
Demenba-small $(3 \rightarrow 2)$	29m	SD	CE	0.82
Demenba-medium	48m	VAD	CE	0.82
		SD	CE	**0.87**
3-class Classification				
EfficientNet b6	40m	VAD	CE	0.62
		SD	CE	0.69
Demenba-small	29m	VAD	CE	0.68
		SD	CE	<u>0.81</u>
Demenba-medium	48m	VAD	CE	0.75
		SD	CE	**0.83**

Demenba, whereas BCE performs better for EfficientNet. In the 3-class setting, we employ a weighted CE loss with weights $(1, 3, 3)$ (normal, MCI, dementia), which yields the highest validation performance. We train each Demenba model for 40 epochs with Adam [39], using an initial learning rate of 10^{-5}, $\beta_1 = 0.95$, $\beta_2 = 0.999$, weight decay $= 5 \times 10^{-7}$, and a batch size of 1 (to fit the longest segments). We warm up for 1,000 steps, then apply an exponential decay of 0.5 per epoch starting at epoch 10. All training is done on a single NVIDIA A6000 GPU (48 GB). We measure performance using the Area Under the Receiver Operating Characteristic Curve (AUC), which enables comparison across all detection thresholds. Results from the best models on the test set are reported, including those on the eval set.

B. Overall results

Table II summarizes our audio-only ADC performance. In the 2-class setting, Demenba-small achieves an AUC comparable to or better than the EfficientNet baseline with significantly fewer parameters. Demenba-medium surpasses EfficientNet by 6% and 5% absolute AUC using VAD and SD boundaries respectively, despite incurring only a modest parameter increase. For 3-class classification, Demenba-small and Demenba-medium improve over EfficientNet by 6% and 13% AUCs using VAD boundaries and 12% and 14% AUCs using SD boundaries respectively. This gap grows as the task becomes more fine-grained, underscoring the benefit of SSMs when discriminating subtle differences between MCI and dementia. Among the Demenba classifier, medium performs better than small by 10% absolute AUC for 2-class and 7% absolute for 3-class, demonstrating the scalability of our method. We also compare two segmentation strategies: a coarse-grained VAD versus a fine-grained SD. Across all models and both classification settings, SD-based boundaries outperform VAD by as much as 5% absolute AUC. Note that we control irrelevant variables such as the segment length to ensure the difference is not due to change in segment length. Instead, SD preserves the within-speaker conversational context (turn-taking, prosodic continuity), which appears crucial for ADC. Lastly, we observed that direct 2-class finetuning outperforms training a 3-class model and merging MCI+dementia outputs for 2-class evaluation by 3 AUC points.

TABLE III: **Dementia classification results for text-only and audio+text models.** X+BERT models use the fully finetuned bert-base-cased (with an MLP head) as the text classifier. X+Llama models use the finetuned Llama-3.1-8B-Instruct as the text-based classifier and X+Qwen2 models use the finetuned Qwen2-7B-Instruct as the text classifier. All audio models use segment boundaries from an SD. "Segment length=Full" means the whole the recording is fed into the model in one forward pass.

	# Trainable Param.	Finetuning Method	Segment Length (s)	AUC (↑)
2-class Classification				
bert-base-cased	109m	Full	180	<u>0.91</u>
Llama-3.1-8B-Instruct	0	No	Full	0.73
	0	No	360	0.53
	21m	LoRA	360	0.83
Qwen2-7B-Instruct	0	No	Full	0.60
	0	No	360	0.80
	20m	LoRA	360	0.85
phi-4	0	No	Full	0.64
	0	No	360	0.65
Demenba-medium+BERT	157m	Full	180	**0.95**
Demenba-medium+Llama	69m	LoRA	360	0.90
Demenba-medium+Qwen2	68m	LoRA	360	0.87
Demenba-small+Llama	68m	LoRA	360	0.87
Demenba-small+Qwen2	68m	LoRA	360	0.85

TABLE IV: **Dementia classification results on the eval set using the best checkpoints from the test set.** X+BERT models use the fully finetuned bert-base-cased (with an MLP head) as the text-based classifier. "m" stands for million and "b" stands for billion. All models use ground truth speaker segment boundaries. $\sim$ denotes average length and without $\sim$ denotes maximal length.

	# Trainable Param.	Input Type	Segment Length (s)	AUC (↑)
2-class Classification				
EfficientNet [6]	-	Audio	$\sim$8	0.78
EfficientNet+LM [6]	-	Audio+Text	$\sim$8	0.83
bert-base-cased	109m	Text	180	0.88
		ASR text	180	0.75
Demenba-medium	48m	Audio	360	**0.81**
Demenba-small	29m	Audio	360	0.71
Demenba-medium+BERT	157m	Audio+Text	360	**0.92**

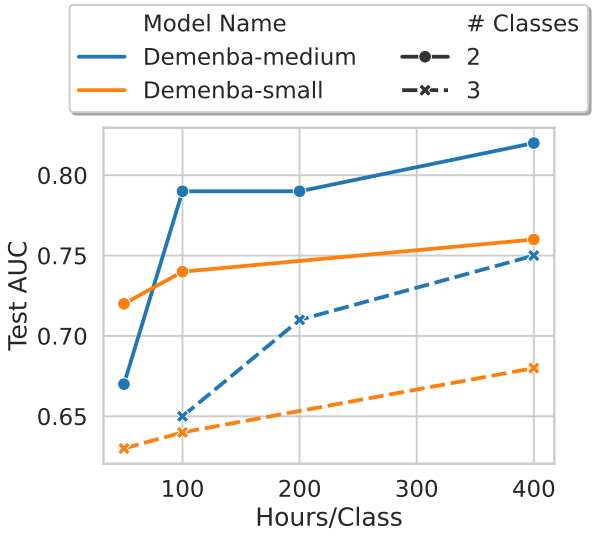
Table III presents performance for text-only and audio+text hybrids. In the 2-class setting, bert-base-cased achieves an AUC of 0.91 — outperforming the best finetuned LLM (Qwen2-7B-Instruct) by 6% absolute. Although recent LLMs show strong general language understanding, most of them perform poorly zero-shot for ADC, likely because of the high

error rate of the ASR transcripts. One exception is Qwen2-7B-Instruct, which achieves an AUC of 0.80 in the zero-shot setting, showing that the LLM has some level of understanding about dementia. Performing fine-tuning on LLMs further narrows the gap but does not surpass BERT, which suggests that, under noisy transcript conditions, a mid-sized PLM fine-tuned end-to-end remains the best text approach for ADC. Importantly, combining modalities yields further gains: the hybrid of BERT+Demenba-medium achieves an AUC of 0.95, the single best result across all systems. This synergy indicates that audio-based and text-based classifiers extract complementary features: audio models capture prosodic cues such as hesitation and intonation, whereas text models capture lexical/linguistic patterns like filler words and semantic incoherence.

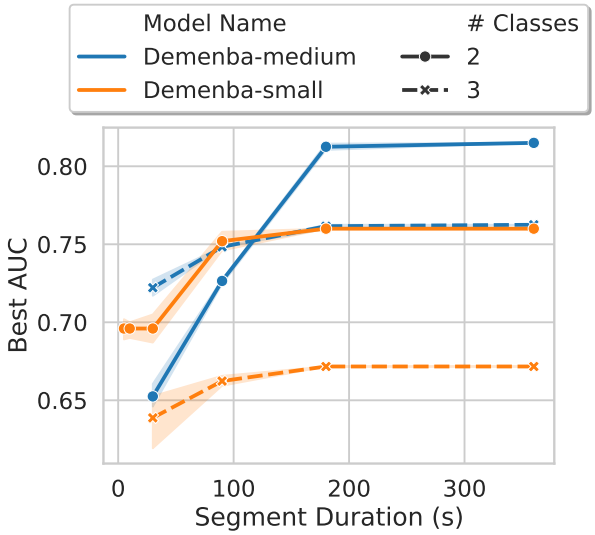
Table IV reports evaluation on the 92-recording dataset with manual transcripts and manual diarization. We benchmark our audio-only Demenba-medium and audio+text Demenba-medium+BERT against EfficientNet and EfficientNet+LM. On this subset, Demenba-medium alone outperforms EfficientNet by 3% absolute AUC, and Demenba-medium+BERT outperforms EfficientNet+LM by 9% absolute AUC, demonstrating the advantage of leveraging longer context. We also found that BERT-based text classifier performs better using real text than using ASR text by 13% absolute AUC, despite being trained on ASR text. One possible reason is that the eval set contains shorter and noisier recordings, resulting in domain mismatch between training and evaluation. The larger AUC gap between Demenba-small and -medium also suggests Demenba to be more generalizable and robust to change in recording conditions.

C. Scaling behavior: sample size and sequence length

Fig. 2 illustrates how Demenba performance evolves as we vary (a) the amount of training data and (b) the maximum segment duration. In Fig. 2a, we plot AUC as a function of total training hours per class, from 50 to 400 hours, for both 2-class and 3-class ADC. Both Demenba-small and Demenba-medium show steadily increasing AUC up to 400 hours per class. Even at the largest data point we tested, the curves have not plateaued for Demenba-medium, suggesting that additional hours would likely yield further gains. The slope of AUC versus hours is steeper for 3-class ADC than 2-class ADC across both model sizes. This aligns with recent scaling-law observations in sequence modeling, where more complex tasks demand more data to resolve finer distinctions [40]. At each data level, Demenba-medium outperforms Demenba-small, and the gap widens as we increase training hours. Further, Demenba-small exhibits diminishing returns after 200 hours per class, whereas Demenba-medium continues improving beyond 400 hours/class. The difference suggests that Demenba-medium better exploits additional examples rather than merely scaling its parameter count. Fig. 2b shows the relationship between maximum segment duration (from 30s to 360s) and AUC. Moving from 30s to 180s yields consistent AUC gains (e.g., 8-17% absolute improvement in 2-class, 4-5% in 3-class). This supports the hypothesis that



(a) Classification AUC vs. training sample size



(b) Classification AUC vs. input segment length

Fig. 2: **Scaling behavior of Demenba with sample size and segment duration.** (a) Scaling behavior with training sample size. All models take 360-second segments as inputs; (b) Scaling behavior with test segment duration. All models are trained on the 400-hour/class datasets. Best AUC is the highest AUC for all segment durations up to the current value. dementia markers (hesitations, prosodic changes) often span multiple turns. However, diminishing return occurs beyond 180s: going from 180s to 360s yields only a marginal 1-2 point AUC increase. Across both variants, 2-class AUC grows more sharply with segment duration than 3-class AUC. Demenba-medium's AUC increases by 17% between 30s and 360s (2-class), whereas Demenba-small improves by 8%. This suggests that Demenba-medium's deeper architecture better captures long-range dependencies.

D. Effect of segment types

We next analyze how different segment types — silence vs. non-silence, and interviewer vs. participant speech —

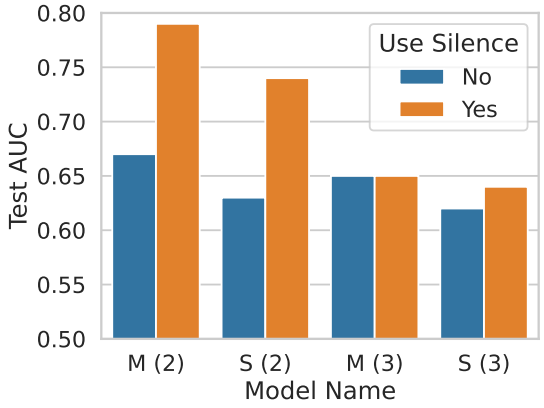


Fig. 3: **Effect of including silences in audio on dementia classification performance for different model sizes and number of classes.** All models are trained with 100-hour/class subset and 360s segments. M (k) and S (k) stand for Demenba-medium (k-class) and Demenba-small (k-class) respectively.

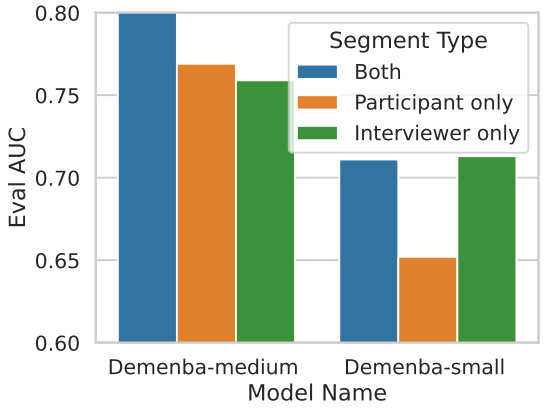


Fig. 4: **Effect of including different speaker role segments for 2-class classification on the eval set.** All models are trained with 360-second segments.

affect ADC performance. Fig. 3 compares AUC when training Demenba on only non-silence frames versus combined speech+silence. Including silence segments yields a 10-15% AUC boost over speech-only models for the 2-class setting, probably because silence often signals hesitation and word-finding difficulty — hallmarks of cognitive decline. In the 3-class setting, adding silence improves AUC by only 0-2% over speech-only baselines. This suggests that while extreme silences help to discriminate "normal" from "advanced dementia", they are less helpful to distinguish between MCI and dementia due to similar hesitation patterns. Fig. 4 examines how model performance varies when training on participant segments only, interviewer segments only, or both. We use speaker boundaries derived from our SD system during training and gold boundaries to isolate each role's speech at test time. Training with both interviewer and participant speech yields the highest AUC of 0.81 for Demenba-medium, about 4% and 5% higher than training with participant speech only and interviewer speech only respectively. On the other hand,

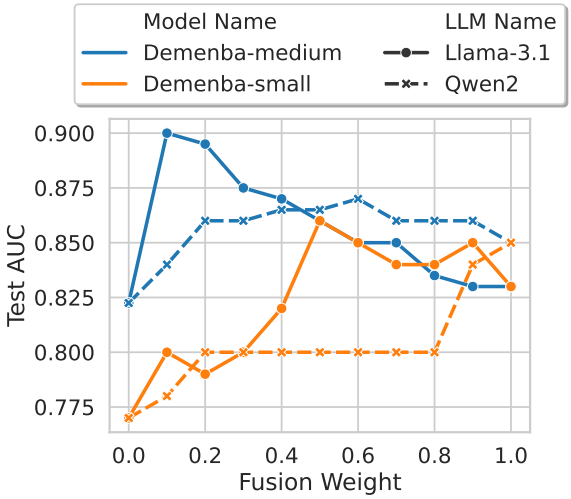


Fig. 5: **Dementia classification performance vs. audio-text fusion weight for various audio and text models.** All models are trained with 360-second segments.

Demenba-small relies mostly on the interviewer segments, as even training on interviewer speech alone yields similar performance to training with both, and training on participant speech alone degrades the AUC by 6%. In general, we found a significant amount of dementia-related information in the interviewer speech, consistent with prior works [6]–[8] on unconscious interviewer bias toward impaired participants.

E. Relative importance of linguistic vs. acoustic information

Fig. 5 examines how AUC varies as we sweep the fusion weight between the audio-only Demenba and the text-only classifier, including LLaMA and Qwen2-7B-Instruct. We observe that the optimal fusion weight to be model-dependent. For Qwen2-7B-Instruct (2-class), the peak AUC occurs at $\lambda \approx 0.5$ (50% reliance on text cues), while for Llama, the peak occurs at $\lambda \approx 0.1$. For both variants of Demenba, we observe a higher fusion weight or higher reliance on the text-based model for Qwen2-7B-Instruct than for Llama.

V. Conclusion

We study the use of SSMs for ADC from neuropsychological tests. Our method outperforms previous best methods, especially in the more challenging 3-class setting. Furthermore, we show that our approach is scalable to over 1,000 hours of speech data during training and long audio up to 6 minutes during both training and inference. Lastly, our analysis on the model sheds new lights on the role of acoustic features, speaker information, linguistic information and reasoning in the ADC task. In the future, we would like to further improve the performance and interpretability of our models by integrating it with multimodal LLMs. Other directions include studying more fine-grained classification of dementia subtypes and generalization performance on other dementia datasets, including those with de-identified speech recordings.

REFERENCES

[1] C. A. Szekely, J. E. Thorne, P. P. Zandi, M. Ek, E. Messias, J. C. Breitner, and S. N. Goodman, "Nonsteroidal antiinflammatory drugs for the prevention of alzheimer's disease: a systematic review," *Neuroepidemiology*, vol. 23, no. 4, pp. 159–169, 2004.

[2] Y. F. Chuang, Y. An, M. Bilgel, D. F. Wong, J. C. Troncoso, R. J. O'Brien, J. C. Breitner, L. Ferrucci, S. M. Resnick, and M. Thambisetty, "Midlife adiposity predicts earlier onset of alzheimer's dementia, neuropathology and presymptomatic cerebral amyloid accumulation," *Molecular Psychiatry*, vol. 21, pp. 910–915, 2016.

[3] L. Kurlowicz and M. Wallace, "The mini-mental state examination (mmse)," *Journal of Gerontological Nursing*, vol. 25, no. 5, pp. 8–9, 1999.

[4] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo challenge," in *Interspeech*, pp. 3780–3784, 2021.

[5] T. Alhanai, R. Au, and J. Glass, "Spoken language biomarkers for detecting cognitive impairment," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 409–416, 2017.

[6] N. Dawalatabad, Y. Gong, S. Khurana, R. Au, and J. Glass, "Detecting dementia from long neuropsychological interviews," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, (Abu Dhabi, United Arab Emirates), pp. 5270–5283, Association for Computational Linguistics, 2022.

[7] T. Al Hanai, R. Au, and J. Glass, "Role-specific language models for processing recorded neuropsychological exams," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (M. Walker, H. Ji, and A. Stent, eds.), (New Orleans, Louisiana), pp. 746–752, Association for Computational Linguistics, June 2018.

[8] P. Pérez-Toro, S. Bayerl, T. Arias-Vergara, J. Vásquez-Correa, P. Klumpp, M. Schuster, E. Nöth, J. Orozco-Arroyave, and K. Riedhammer, "Influence of the interviewer on the automatic assessment of alzheimer's disease in the context of the adresso challenge," in *Proceedings of Interspeech*, pp. 3785–3789, 2021.

[9] Vaswani *et al.*, "Attention is all you need," in *NeurIPS*, p. 6000–6010, 2017.

[10] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[11] A. Balagopalan and J. Novikova, "Comparing acoustic-based approaches for alzheimer's disease detection," in *Interspeech*, pp. 3800–3804, 2021.

[12] J. Li, K. Song, J. Li, B. Zheng, D. Li, X. Wu, X. Liu, and H. Meng, "Leveraging pretrained representations with task-related keywords for alzheimer's disease detection," in *ArXiv*, 2023.

[13] S. Bhati, Y. Gong, L. Karlinsky, H. Kuehne, R. Feris, and J. Glass, "Dass: Distilled audio state space models are stronger and more duration-scalable learners," in *SLT*, 2024.

[14] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *ICLR*, 2022.

[15] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[16] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Interspeech*, (Shanghai, China), 2020.

[17] M. Rohanian, J. Hough, and M. Purver, "Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs," in *Interspeech*, pp. 3820–3824, 2021.

[18] C. Xue, C. Karjadi, I. C. Paschalidis, R. Au, and V. B. Kolachalama, "Detection of dementia on voice recordings using deep learning: a framingham heart study," *Alzheimer's Research & Therapy*, vol. 13, p. 146, Aug. 2021.

[19] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[21] R. Haulcy and J. Glass, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, p. 624137, 2021.

[22] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[23] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[24] M. H. Erol, A. Senocak, J. Feng, and J. S. Chung, "Audio mamba: Bidirectional state space model for audio representation learning," *IEEE Signal Processing Letters*, vol. 31, pp. 2975–2979, 2024.

[25] J. Lin and H. Hu, "Audio mamba: Pretrained audio state space model for audio tagging," *arXiv preprint arXiv:2405.13636*, 2024.

[26] S. Shams, S. S. Dindar, X. Jiang, and N. Mesgarani, "Ssamba: Self-supervised audio representation learning with mamba state space model," *arXiv preprint arXiv:2405.11831*, 2024.

[27] X. Zhang, Q. Zhang, H. Liu, T. Xiao, X. Qian, B. Ahmed, E. Ambikairajah, H. Li, and J. Epps, "Mamba in speech: Towards an alternative to self-attention," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[28] X. Jiang, Y. A. Li, A. N. Florea, C. Han, and N. Mesgarani, "Speech slytherin: Examining the performance and efficiency of mamba for speech separation, recognition, and synthesis," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2025.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, Association for Computational Linguistics, 2019.

[30] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. INTERSPEECH 2023*, 2023.

[31] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. INTERSPEECH 2023*, 2023.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, IEEE, 2009.

[33] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (New Orleans, LA, USA), IEEE, 2017.

[34] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, vol. 97, pp. 6105–6114, PMLR, 2019.

[35] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lanchaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[36] A. Yang *et al.*, "Qwen2 technical report," tech. rep., Qwen Team, Alibaba Group, 2024. Technical Report.

[37] Microsoft, "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," *Technical Report*, 2024.

[38] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, (San Diego, CA), 2015.

[40] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.