Early Detection of Alzheimer's Disease from Speech with Deep Learning

Master's Thesis Progress Report

Raftopoulos Michail

National Technical University of Athens, Greece

February 27, 2026

Latest additions can be found in Section 2: Dataset and Section 3: Methodology.

Abstract

Early detection of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) is critical for timely intervention and effective disease management. Speech analysis has emerged as a promising, non-invasive, and cost-effective approach for detecting the cognitive decline associated with AD and MCI. This master's thesis aims to contribute to the field by focusing on modeling the conversational nature of the available datasets, a crucial yet often overlooked characteristic. In this report, we document our progress through the stages of data analysis and preprocessing, baseline experiments and experimentation regarding conversational modeling, deriving inspiration from the field of paralinguistics.

1 Introduction

Dementia is a progressive neurodegenerative disorder that causes irreversible brain damage, significantly impairing activities of daily living. It is estimated to affect over 57.4 million people, a figure that is expected to rise dramatically due to the aging population [1]. Alzheimer's Disease (AD) is the most prevalent form of dementia, accounting for 60-70% of all cases [2].

While there is no known cure for AD, early detection can greatly help patients manage their symptoms and improve the quality of life for them and their families. In particular, researchers highlight the need for detecting a cognitive decline stage known as Mild Cognitive Impairment (MCI) which may or may not progress to dementia. Speech analysis has emerged as a promising approach for detecting MCI, offering a non-invasive, cost-effective, and accessible means of assessment [3], [4].

The field has witnessed significant academic interest in recent years, characterized by frequent competitions and a growing body of literature. However, it continues to face substantial challenges that hinder clinical translation. Although many models report high evaluation metrics, performance levels remain insufficient for reliable clinical application. Moreover, these models are frequently trained and evaluated on small or outdated datasets, raising concerns regarding their generalizability. In this research proposal, we perform a comprehensive literature review of existing methodologies and the evaluation metrics achieved. We then identify gaps in the literature and missed opportunities, and finally propose a set of promising methodologies to address these limitations.

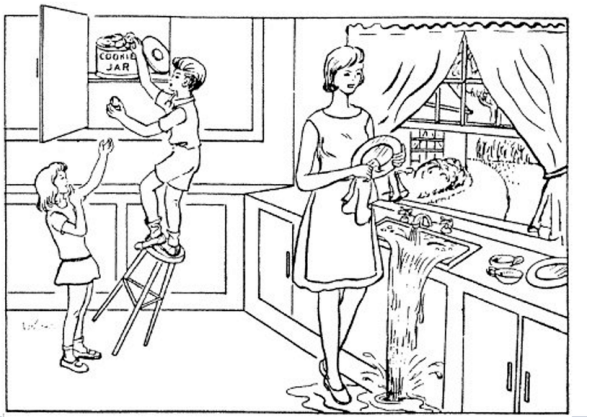

Figure 1: The Cookie Theft Picture, used in the picture description task of the Pitt corpus and other datasets from DementiaBank.

2 Dataset

Most available data in the field is provided by the *DementiaBank* [4] database. DementiaBank consists of a collection of corpora, that contain audio recordings and CHAT transcripts of various cognitive assessment tasks, administered by an interviewer[1]. In this work, we utilize the *MultiConAD* [5] framework, which aggregates all DementiaBank corpora into a single, unified dataset. In addition, MultiConAD converts all transcripts to plain text and experiments with translation and multilingual trainings. We choose to focus only on the purely English subset of MultiConAD.

Implementation Details: While the MultiConAD reproduction source code is available online[2], it was realised that the repository was missing critical preprocessing steps that were essential for the incorporation of the "WLS", "VAS" and "Kempler" corpora. Thus, these scripts had to be manually implemented. In addition, since the publication of the MultiConAD dataset, the "Delaware" corpus has grown in number of data samples, that we chose to incorporate into our implementation. As a result, our MultiConAD implementation differs slightly from the original one. The differences between the two implementations are summarized in Table 1.

MultiConAD	Tot Length	Samples	Train	Test
Original	*not specified*	2411	2201	210
Ours	114.49 hours	2469	2217	252

Table 1: Differences in data samples between ours and the original (English only) MultiConAD dataset implementation.

After the implementation and analysis of the dataset at hand, several key observations have emerged that will shape the direction of this study's methodology:

1. **Conversational Data**: As already mentioned, most data samples consist of oral conversations. To our knowledge, no recent study has modeled this conversational structure, leading to models processing the speech of both the interviewer and the participant with no distinction whatsoever (see 5.3.3 in the Literature Review). This is where choose to focus our study.

2. **Small Dataset**: Although it includes nearly all available datasets, MultiConAD is still severely small for deep learning standards, amounting to about 2500 samples, or about 115 hours. Thus, the resulting model will be limited to fine-tuning, or to utilizing layers of open, pre-trained language and paralinguistic models.

3. **Class Imbalance**: As shown in Figure 2 (top and bottom right), the class distribution heavily favors the healthy controls. This raises the need for class-balancing data augmentation.

4. **Large-Scale Recordings**: The DementiaBank dataset consists of substantial audio files with significant variance in duration, ranging from 2 to 35 minutes. Given that most speech models are optimized for short windows of a few seconds, we have introduced a design requirement to manage the aggregation of these processed segments into a cohesive output. This can be better demonstrated by Figure 2 (top left).

5. **Dataset Bias**: The individual sub-datasets within MultiConAD exhibit a disproportionate class distribution (see Figure 2, bottom left). For instance, the WLS dataset comprises a vast majority of healthy controls. This imbalance introduces a significant bias hazard; the model may inadvertently learn to classify samples based on the unique acoustic signatures or noise fragments of the source datasets rather than the underlying pathological voice characteristics.

3 Methodology

In this section we document our progress regarding baselines, candidate options for further experiments, and the results yielded by our models.

3.1 Baseline Experiments

The original MultiConAD paper proposes a series of simple baseline models that utilize the transcripts' text representations, in conjunction with traditional ML classifiers. For text representations, the authors experimented with TF-IDF representations (sparse) and `intfloat/multilingual-e5-large` embeddings (dense). Since our implementation of MultiConAD differs slightly from the original, we chose to reproduce these benchmarks to ensure a fair comparison. The results can be seen in Table 2.

Model	Acc	UAR	F1
MultiConAD - Sparse (original)	0.65	-	-
MultiConAD - Dense (original)	0.65	-	-
MultiConAD - Sparse (ours)	0.7	0.66	0.68
MultiConAD - Dense (ours)	0.67	0.63	0.65

Table 2: Baseline model evaluation metrics.

[1] The only exception is the VAS corpus, which consists of recorded Voice Assistant System commands.
[2] https://github.com/ArezoShakeri/MultiConAD

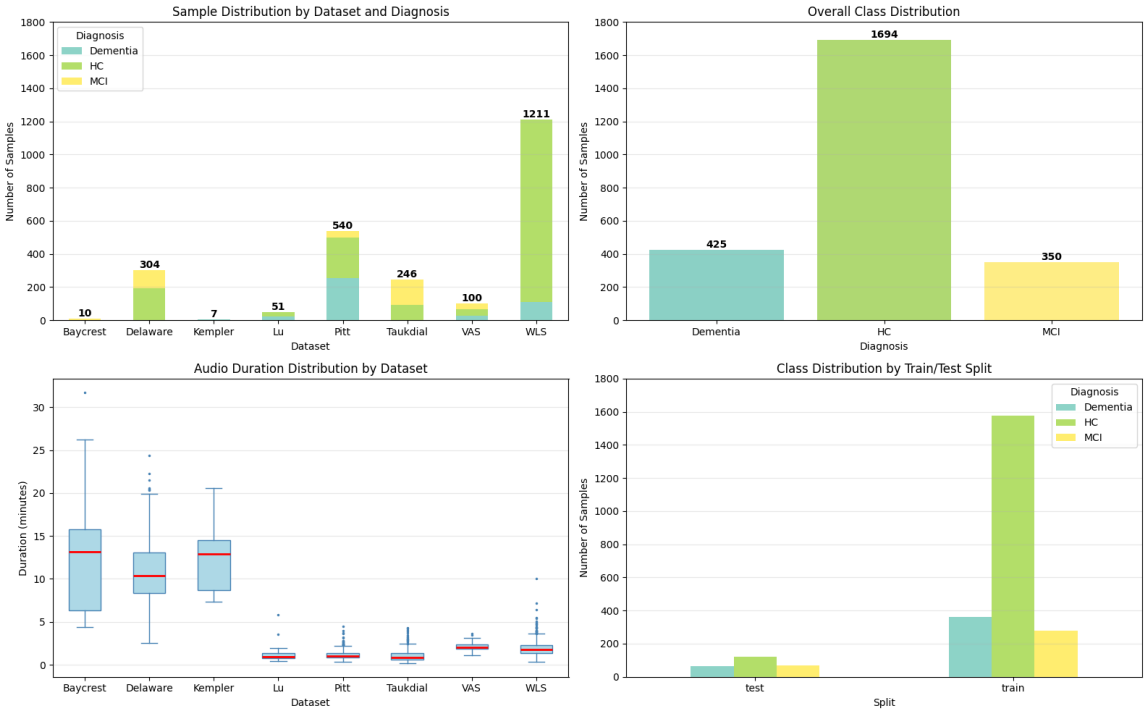

Figure 2: *top left:* Sample and class distribution among sub-datasets of MultiConAD. *bottom left:* Recording durations among sub-datasets of MultiConAD. *top right:* Class distribution. *bottom right:* Class distribution by Train/Test split.

I'm not sure why the accuracy gap between ours and the original MultiConAD implementation is so significant.

I could try to reproduce more papers on MultiConAD but since it is very cumbersome to work with, I'd like to move forward with the methodology and perhaps return to that later.

3.2 Conversational Modeling

This study focuses on modeling the conversational structure of DementiaBank data, to assist the model in differentiating the interviewer from the participant, control each person's involvement in the model's decision-making and evaluate the participant's question answering skills in a more precise way.

Deriving inspiration from the field of Multimodal Emotion Recognition in Conversations [6]–[8], we distinguish two possible directions for our methodology:

- **Prompt-based**: Having a foundational language model as a decision-making backbone, we engineer prompts to incorporate speaker text, speaker information, and/or speech embeddings.
- **Graph-based**: We model the conversation graph, with nodes representing speaker utterances and edges representing semantic connections.

[3] https://researchers.wls.wisc.edu/

Keeping the limited dataset size in mind, we need to restrict the number of trainable parameters and base the model on pretrained layers. In future steps, we will focus on one of these options (at least initially) and flesh out an initial design, from where the experimentations will begin.

4 Future Steps

- Perform initial experiments with one of the two mentioned methodologies.
- Reproduce more models on the MultiConAD dataset.

5 Notes

- Although MultiConAD provides 1211 artificially labeled WLS recordings, the total amount of WLS samples provided by DementiaBank amount to approximately 10000 recordings from about 1300 speakers. Moreover, even more data samples may be retrieved by contacting the official WLS study organization[3]. This opens the door to potential self-supervised or silver-labeling techniques as a means to mitigate the limited dataset size issue.

Appendix: Literature Review

5.1 Datasets and Evaluation Challenges

DementiaBank: Most datasets currently available are provided by the DementiaBank database. DementiaBank is a large collection of audio recordings, consisting of various cognitive assessment tasks, administered by an interviewer. Most of these recordings are manually transcribed in the CHAT format, a transcription method that incorporates various linguistic observations such as pauses and stutters, speaker roles, and timestamps along with the spoken words. The most widely used of the provided datasets is the Pitt corpus, which includes mostly the "Cookie Theft Picture Description" task, where participants are asked to describe the image shown in Figure 1. There has also been work toward the creation and expansion of the Delaware corpus, a new dataset that focuses on the binary MCI vs HC distinction. It provides a large variety of cognitive tasks and includes subjects of a wide ethnic and cultural diversity [4].

WLS: The WLS is a large-scale, extended longitudinal study of a random sample of 10,317 men and women who graduated from Wisconsin high schools in 1957. The WLS participants were interviewed up to 6 times between 1957 and 2011. DementiaBank provides access to a subset of audio recordings from the 2003 and 2011 interview rounds, amounting to about 1300 different speakers. Although this is a dataset of substantial size, it lacks clinical labels for cognitive impairment, limiting its direct applicability to supervised learning tasks.

ADReSS, ADReSSo and ADReSS-M: The ADReSS Challenge was introduced in the Interspeech 2020 conference. It provides a balanced subset of the Pitt corpus, with respect to age and gender [9]. The ADReSSo Challenge followed in 2021, introducing a more difficult task of AD detection using only speech samples, without manually created transcriptions. It also utilized a subset of the Pitt corpus [10]. Lastly, the ADReSS-M challenge, introduced in the ICASSP 2023 conference, focused on the binary classification task of AD detection in the Greek language. It provided a subset of the Pitt corpus, as well as a new dataset with Greek speech samples [11]. Even after the end of these challenges, the datasets remain available and have been widely used as benchmarks for new approaches.

MultiConAD: The MultiConAD dataset [5] is a recent effort to tackle the problems of 3-class classification and multilingual generalization. It combines multiple existing datasets, mostly the ones provided by DementiaBank, to create a large and diverse multilingual dataset, with a variety of cognitive assessment tasks. It includes audio and transcription samples in English, Spanish, Chinese, and Greek. Additionally, it provides an important set of baseline models that act as a starting point for future research.

PROCESS: The ICASSP 2025 PROCESS Grand Challenge [12] introduced a modern dataset to serve as a benchmark for the 3-class detection problem. It provides audio recordings and manual transcripts (with only audio provided for the test set) from three cognitive tasks: Semantic Fluency, Phonemic Fluency, and Picture Description. The contestants were also tasked with performing regression on the MMSE score, a widely used clinical assessment for cognitive impairment.

5.2 AD and MCI Detection from Speech

In this section, we summarize the general trends in methodologies tackling AD and MCI detection from speech. We organize our review by general methodologies; however, these subsections loosely resemble the chronological development of approaches.

5.2.1 Feature Engineering

Initial work emphasized the extraction of information-rich features, used in conjunction with traditional ML methods. These approaches proved very effective in the challenge settings as well, where the applications were limited to the small challenge datasets. These methods extract mostly hand-crafted acoustic (e.g., pauses, low-level descriptors, etc.) and linguistic (e.g., verbal richness, filler words, etc.) features and/or semantic representations, mostly from the BERT [13] language model.

Such approaches have led to winning submissions in the ADReSS and ADReSSo Challenges [14], [15]. In the more recent PROCESS Challenge, the overall winner leveraged linguistic features, extracted from ASR transcriptions, and an ensemble of traditional machine learning models, to achieve an F1 score of 0.649. These linguistic features include cognitive-task-specific indicators (count of correct words in Verbal Fluency Tasks) and algorithmically extracted speech pause descriptors [16]. The best performing submission specific to the classification task achieved an F1 score of 0.696, using a self-developed Digital Linguistic Biomarker (DLB) extractor [17].

In a notable paradigm of this methodology, the authors of ADscreen [18] extract a broad range of features, including metrics related to phonetic motor planning, semantic disfluency (e.g., word repetition and pausing), lexical diversity, and syntactic structure. They further incorporate BERT embeddings to capture verbal disfluencies, as well as psycholinguistic features such as LIWC and GeMAPS. Similarly, Haider et al. focus on the emotional dimension of speech by introducing the Affective Behaviour Representation (ABR), which employs a machine learning model to label each speech segment with an emotion and summarize the emotional content of an entire recording into a single vector [19].

5.2.2 Deep Learning Architectures and Multimodal Fusion

While feature engineering relies on extracting predefined markers, deep learning approaches aim to learn latent representations directly from the data. Given the success of deep learning in the more general field of paralinguistics, researchers naturally sought to implement it in the AD detection from speech domain. Deep learning architectures were used to produce meaningful latent representations and fuse acoustic and textual cues.

Within early works, Liu et al. [20] feed MFCCs into a DNN encoder to produce low-dimension bottleneck features of 40ms time frames, followed by CNN and BiLSTM layers. Evaluating their model with 10-fold CV on the Pitt corpus, they achieved an accuracy of 82.59% and an F1 of 82.94%.

In the multimodal setting, a wide variety of methods and architectures emerged. These approaches experiment with various combinations of features, such as BERT embeddings, traditional acoustic features, and other neural-network extracted features [21]–[25]. In an interesting approach, Lee et al. achieved state-of-the-art results in the ADReSSo Dataset by processing the Cookie Theft Picture description task image with VLMs and comparing it with the patient's text modality to check the validity of the response. They additionally utilized the Shapley value from game theory to introduce a new auxiliary loss function, which includes information about each modality's contribution [26].

5.2.3 Fine-tuning and Foundational Models

Foundational models have been influencing the domain of AD detection from speech since the early days of the ADReSS Challenge. As mentioned in the preceding paragraphs, BERT was frequently used to extract semantic embeddings that were later used as features for classification. Today, Large Language Models are still widely used [27]–[29], utilizing ASR transcriptions from state-of-the-art models and leading to promising results. An extensive study that compared various text-based ML methods showed that the fine-tuning of pretrained models is the most performant approach [30], at least in the text-only setting.

However, limiting the input to text only discards very crucial biomarkers embedded in the patient's voice. While some have resorted to encoding audio information into the transcripts [27], others have fine-tuned the ASR models themselves to allow the processing of purely acoustic inputs [31]–[33].

Most recent works merge acoustic and linguistic information by incorporating the use of Large Audio-Language Models [34], [35]. Zolnour et al. use a late fusion architecture, combining the predictions of a classifier that is fed pretrained encoder features, and a classifier that is fed handcrafted linguistic features. They also perform extensive experimentations with unimodal and multimodal LLMs, fine-tuned on classification from text and audio. Lastly, they generate synthetic text by prompting foundational LLMs. Overall, the late fusion scheme outperformed the fine-tuned LLMs [34]. Shanin et al. bypass fine-tuning entirely by simply prompting an Audio-Language model. Although the results were not optimal, they were comparable with supervised methods [35].

Lastly, regarding intermediate pretraining, Zhu et al. [36] used language model perplexity metrics to select large datasets from the GLUE benchmark that are maximally similar to the ADReSS Challenge dataset. Then, they pre-trained a language model on the chosen datasets, followed by fine-tuning on the AD detection task. They additionally utilized their used perplexity-based metrics to invent a new sample-level pretraining technique, where samples that don't reduce perplexity are discarded.

5.2.4 Data Augmentation

Despite the known issue of data scarcity in the field, the applications of data augmentation are limited. To our knowledge, only two publications in the field study explicitly the creation of synthetic data. The authors of CDA [37] propose a contrastive data augmentation technique that simulates cognitive decline by removing random words from sentences, and generates positive samples with multiple passes in conjunction with dropout. Hlédiková et al. [38] perform extensive experimentation with various data-space augmentation techniques. The methods included classical acoustic and verbal perturbations and deep learning-based ones, namely voice conversion with the FragmentVC model, lexical paraphrasing with the Pegasus model and text generation using GPT-2. The authors conclude that their tested neural-based methods perform similarly to traditional ones, still achieving significantly high results.

Some other works have incorporated data augmentation into their methodologies. Specifically, Liu et al. [20] utilize SpecAugment [39], Runde et al. [29] apply Synthetic Minority Over-sampling to balance out the datasets, and Lin and Washington perform Synonym Replacement for their text-based model. Additionally, the authors of LLMCARE [34] experiment with various LLMs to generate synthetic transcripts.

5.2.5 State Of The Art

In Table 3, we summarize the datasets that serve as benchmarks in the field, detailing their sample sizes and number of classes. Alongside each dataset, we report the most successful model we could locate in the literature and its corresponding performance metrics. However, we emphasize that the literature domain is currently quite fragmented. The standardized contest datasets are

limited in size, and larger alternatives often lack determined train-test splits or standardized evaluation methods (e.g., simple inference vs. cross-validation). This lack of standardization makes objective comparisons and the identification of a true SOTA challenging. Despite these limitations, this overview highlights the primary data resources available and the current state of performance.

All mentioned dataset, besides the Pitt Corpus and MultiConAD were introduced as evaluation challenges. ADReSS and ADReSSo are subsets of the Pitt Corpus. MultiConAD is an aggregated dataset combining multiple DementiaBank corpora. Although it is multilingual, we focus only on the English subset. Since MultiConAD is a very recent dataset, the only model for comparison is the baseline set by the authors.

5.3 Limitations and Opportunities

5.3.1 Generalizability Concerns

While the scientific community has achieved remarkable evaluation metrics and made impressive strides in the field, these advancements are currently limited by the small, often outdated and noisy available datasets—often containing only a few hundred samples. Consequently, the applicability of these models in a clinical setting must be questioned.

In a study by Runde et al. [29], the authors managed to achieve 0.99 accuracy and F1 on the Pitt Corpus - a superset of the ADReSS and ADReSSo datasets - by utilizing Wav2Vec transcripts and ada-002 text embeddings. However, when tested on a 10-fold cross validation scheme, these scores dropped to 0.79. A more concerning example involves researchers achieving nearly 100% accuracy on the Pitt Corpus, using solely the silent segments of the audio recordings. This demonstrates the presence of a "Clever Hans" effect, where models achieve high accuracy not by actually learning the underlying mechanisms of the problem, but through spurious correlations in the training data [45]. These findings strongly emphasize the need for a larger, more diverse dataset, a gap promised to be filled by MultiConAD.

5.3.2 Interpretability

Clinical adoption necessitates model interpretability, as healthcare professionals must understand which speech characteristics drive diagnostic predictions in order to integrate findings with other clinical indicators and establish professional trust. However, most current approaches utilize complex deep learning architectures and pre-trained language models that function as "black boxes," thereby obscuring the decision-making process. While the domain may not yet have prioritized interpretability as an immediate prerequisite, future methodologies must be designed with the understanding that clinical viability ultimately depends on the transparency and interpretability of the model.

5.3.3 Untreated Conversational Nature of Data

With the exception of the 100 samples of the VAS (Voice Assistant System) corpus, all of the DementiaBank data consist of recorded dialogues, where an interviewer administers a cognitive test with the patient. To our knowledge, no recent study has accounted for this conversational structure, instead treating the recordings as homogeneous samples. Consequently, models analyze semantic and paralinguistic information from both the patient and the interviewer indiscriminately. Previous studies have highlighted this limitation, demonstrating that the interviewer's speech significantly impacts observed linguistic features, thereby potentially introducing bias and confounding model predictions [46].

5.3.4 Addressing Data Limitations

The most significant constraint within the studied domain is the limited size of available datasets, which restricts methodological innovation and hampers the generalization capabilities of proposed models. Furthermore, the acquisition and dissemination of such data are inherently slow processes due to severe data privacy concerns. Consequently, it is expected that data availability in this field will consistently lag behind industry standards established for more general tasks. While advanced techniques for mitigating data scarcity–such as data augmentation, synthetic data generation, and advanced fine-tuning–should be central to research in this area, they appear to be underrepresented in the current literature. Moreover, the WLS dataset remains significantly underutilized. Despite being unlabeled, this dataset offers substantial potential for pre-training and silver labeling strategies prior to fine-tuning on labeled data. To the best of our knowledge, the only studies that have leveraged WLS in this manner are [5] and [47], which manually determine potential labels based on task performance criteria.

References

[1] E. Nichols, J. D. Steinmetz, S. E. Vollset, *et al.*, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the Global Burden of Disease Study 2019," *The Lancet Public Health*, vol. 7, no. 2, e105–e125, Feb. 2022.

[2] M. V. F. Silva, C. d. M. G. Loures, L. C. V. Alves, L. C. de Souza, K. B. G. Borges, and M. d. G. Carvalho, "Alzheimer's disease: Risk factors and potentially protective measures," *Journal of Biomedical Science*, vol. 26, p. 33, May 2019.

Dataset	Train/Test Split	Classes	Modalities	Performance	Source
Pitt Corpus	552 total (10-fold CV)	2 (AD, HC)	Audio, CHAT Transcripts	**Acc**: 0.95 **F1**: 0.95 **AUR**: 0.93	[40]
ADReSS	Train: 108 Test: 48	2 (AD, HC)	Audio, CHAT Transcripts	**Acc**: 0.94	[41]
ADReSSo	Train: 166 Test: 71	2 (AD, HC)	Audio	**Acc**: 0.96 **F1**: 0.96	[42]
ADReSS-M	Train: 237 Test: 46	2 (AD, HC)	Audio	**Acc**: 0.87 **RMSE**: 3.73	[43]
TAUKADIAL	Train: 387 Test: 120	2 (MCI, HC)	Audio	**UAR**: 0.86	[44]
PROCESS	Train: 157 Test: Hidden	3 (AD, MCI, HC)	Audio, Transcripts	**Macro-F1**: 0.696	[17]
MultiConAD (English subset)	Train: 2201 Test: 210	3 (AD, MCI, HC)	Audio, Transcripts	**Bin Acc**: 0.90 **Ter Acc**: 0.65	[5]

Table 3: State-of-the-Art (SOTA) Models for Alzheimer's Disease Detection from Speech.

[3] "2018 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, Mar. 2018.

[4] A. M. Lanzi, A. K. Saylor, D. Fromm, H. Liu, B. MacWhinney, and M. L. Cohen, "DementiaBank: Theoretical Rationale, Protocol, and Illustrative Analyses," *American Journal of Speech-Language Pathology*, vol. 32, no. 2, pp. 426–438, Mar. 2023.

[5] A. Shakeri, M. Farmanbar, and K. Balog, *MultiConAD: A Unified Multilingual Conversational Dataset for Early Alzheimer's Detection*, Feb. 2025. arXiv: 2502.19208 [cs].

[6] C. Wu, Y. Cai, Y. Liu, *et al.*, *Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects*, Sep. 2025. arXiv: 2505.20511 [cs].

[7] G.-T. Lin, P. G. Shivakumar, A. Gandhe, *et al.*, *Paralinguistics-Enhanced Large Language Modeling of Spoken Dialogue*, Jan. 2024. arXiv: 2312.15316 [cs].

[8] H. Chen, Z. Li, Y. Song, *et al.*, *GOAT-SLM: A Spoken Language Model with Paralinguistic and Speaker Characteristic Awareness*, Jul. 2025. arXiv: 2507.18119 [cs].

[9] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, *Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge*, Aug. 2020. arXiv: 2004.06833 [eess].

[10] S. Luz and F. Haider, "Detecting cognitive decline using speech only: The ADReSSO Challenge,"

[11] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. Macwhinney, "An Overview of the ADReSS-M Signal Processing Grand Challenge on Multilingual Alzheimer's Dementia Recognition Through Spontaneous Speech," *IEEE Open Journal of Signal Processing*, vol. PP, pp. 1–12, Mar. 2024.

[12] F. Tao, B. Mirheidari, M. Pahar, *et al.*, *Early Dementia Detection Using Multiple Spontaneous Speech Prompts: The PROCESS Challenge*, Dec. 2024. arXiv: 2412.15230 [cs].

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, May 2019. arXiv: 1810.04805 [cs].

[14] R. Haulcy and J. Glass, "Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech," *Frontiers in Psychology*, vol. 11, Jan. 2021.

[15] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, "Tackling the ADRESSO Challenge 2021: The MUET-RMIT System for Alzheimer’s Dementia Recognition from Spontaneous Speech," in *Interspeech 2021*, ISCA, Aug. 2021, pp. 3815–3819.

[16] Y. Gao, L. Guo, and H. Liu, *Leveraging Multimodal Methods and Spontaneous Speech for Alzheimer's Disease Identification*, Feb. 2025. arXiv: 2412.09928 [cs].

[17] S. Zhang, N. Khlif, M. Ferro, G. Gagliardi, and F. Tamburini, "Cognitive Decline Detection using DLB Extraction Pipelines," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–2.

[18] M. Zolnoori, A. Zolnour, and M. Topaz, "ADscreen: A Speech Processing-based Screening System for Automatic Identification of Patients with Alzheimer's Disease and Related Dementia," *Artificial intelligence in medicine*, vol. 143, p. 102 624, Sep. 2023.

[19] F. Haider, S. de la Fuente Garcia, P. Albert, and S. Luz, *Affective Speech for Alzheimer's Dementia Recognition*. May 2020.

[20] Z. Liu, Z. Guo, Z. Ling, and Y. Li, "Detecting Alzheimer's Disease from Speech Using Neural Networks with Bottleneck Features and Data Augmentation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 7323–7327.

[21] L. Ilias, D. Askounis, and J. Psarras, "A Multimodal Approach for Dementia Detection from Spontaneous Speech with Tensor Fusion Layer," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Sep. 2022, pp. 1–5. arXiv: 2211.04368 [cs].

[22] L. Ilias and D. Askounis, "Multimodal Deep Learning Models for Detecting Dementia From Speech and Transcripts," *Frontiers in Aging Neuroscience*, vol. 14, Mar. 2022.

[23] Y. Pan, Y. Shi, Y. Zhang, and M. Lu, *Swin-BERT: A Feature Fusion System designed for Speech-based Alzheimer's Dementia Detection*, Oct. 2024. arXiv: 2410.07277 [eess].

[24] J. Cheng, M. Elgaar, N. Vakil, and H. Amiri, *CogniVoice: Multimodal and Multilingual Fusion Networks for Mild Cognitive Impairment Assessment from Spontaneous Speech*, Jul. 2024. arXiv: 2407.13660 [cs].

[25] K. Lin and P. Y. Washington, "Multimodal deep learning for dementia classification using text and audio," *Scientific Reports*, vol. 14, no. 1, p. 13 887, Jun. 2024.

[26] B. Lee, H. J. Song, Y.-J. Park, and B. O. Kang, "Multimodal Alzheimer's disease recognition from image, text and audio," *Scientific Reports*, vol. 15, no. 1, p. 29 038, Aug. 2025.

[27] X. Ke, M.-W. Mak, and H. Meng, "Optimizing Pause Context in Fine-Tuning Pre-trained Large Language Models for Dementia Detection," in *Interspeech 2025*, ISCA, Aug. 2025, pp. 1408–1412.

[28] C. Park, A. S. G. Choi, S. Cho, and C. Kim, "Reasoning-Based Approach with Chain-of-Thought for Alzheimer's Detection Using Speech and Large Language Models,"

[29] B. S. Runde, A. Alapati, and N. G. Bazan, "The Optimization of a Natural Language Processing Approach for the Automatic Detection of Alzheimer's Disease Using GPT Embeddings," *Brain Sciences*, vol. 14, no. 3, p. 211, Feb. 2024.

[30] B. Ihnaini, Y. Deng, Y. He, L. Geng, and J. Xu, "Detection of Alzheimer's Disease Using Fine-Tuned Large Language Models," *Forum for Linguistic Studies*, vol. 7, no. 8, pp. 373–384, Aug. 2025.

[31] K. Jia, J. Li, K. Li, and W.-Q. Zhang, "Whisper-Based Multilingual Alzheimer's Disease Detection and Improvements for Low-Resource Language,"

[32] E. Akinrintoyo, N. Abdelhalim, and N. Salomons, *WhisperD: Dementia Speech Recognition and Filler Word Detection with Whisper*, May 2025. arXiv: 2505.21551 [eess].

[33] J. Li and W.-Q. Zhang, "Whisper-Based Transfer Learning for Alzheimer Disease Classification: Leveraging Speech Segments with Full Transcripts as Prompts," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 11 211–11 215.

[34] A. Zolnour, H. Azadmaleki, Y. Haghbin, *et al.*, "LLMCARE: Early detection of cognitive impairment via transformer models enhanced by LLM-generated synthetic data," *Frontiers in Artificial Intelligence*, vol. 8, Nov. 2025.

[35] M. Shahin, B. Ahmed, and J. Epps, *Zero-Shot Cognitive Impairment Detection from Speech Using AudioLLM*, 2025.

[36] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, "Domain-aware Intermediate Pretraining for Dementia Detection with Limited Data," in *Interspeech 2022*, ISCA, Sep. 2022, pp. 2183–2187.

[37] J. Duan, F. Wei, J. Liu, H. Li, T. Liu, and J. Wang, "CDA: A Contrastive Data Augmentation Method for Alzheimer's Disease Detection," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 1819–1826.

[38] A. Hlédiková, D. Woszczyk, A. Akman, S. Demetriou, and B. Schuller, *Data Augmentation for Dementia Detection in Spoken Language*, Jul. 2022. arXiv: 2206.12879 [cs].

[39] D. S. Park, W. Chan, Y. Zhang, *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech 2019*, Sep. 2019, pp. 2613–2617. arXiv: 1904.08779 [eess].

[40] S. Latif, N. U. Islam, Z. Uddin, K. M. Cheema, S. S. Ahmed, and M. F. Khan, "Deep ensemble learning with transformer models for enhanced Alzheimer's disease detection," *Scientific Reports*, vol. 15, p. 24 720, Jul. 2025.

[41] M. Martinc, F. Haider, S. Pollak, and S. Luz, "Temporal Integration of Text Transcripts and Acoustic Features for Alzheimer's Diagnosis Based on Spontaneous Speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 642 647, Jun. 2021.

[42] N. Ntampakis, K. Diamantaras, I. Chouvarda, M. Tsolaki, V. Argyriou, and P. Sarigianndis, "NeuroXVocal: Detection and Explanation of Alzheimer's Disease through Non-invasive Analysis of Picture-prompted Speech," in vol. 15973, 2026, pp. 410–419. arXiv: 2502.10108 [cs].

[43] X. Chen, Y. Pu, J. Li, and W.-Q. Zhang, "Cross-Lingual Alzheimer's Disease Detection Based on Paralinguistic and Pre-Trained Features," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–2.

[44] M. Bilal, W. Abdulla, G. Cheung, L. Tippett, and S. R. Shahamiri, "Multimodal speech analysis for early detection of mild cognitive impairment: A scalable approach," Oct. 2025, pp. 2465–2470.

[45] Y.-L. Liu, R. Feng, J.-H. Yuan, and Z.-H. Ling, *Clever Hans Effect Found in Automatic Detection of Alzheimer's Disease through Speech*, Jun. 2024. arXiv: 2406.07410 [eess].

[46] C. Li, Z. Sheng, T. Cohen, and S. Pakhomov, *"Is There Anything Else?": Examining Administrator Influence on Linguistic Features from the Cookie Theft Picture Description Cognitive Test*, Mar. 2025. arXiv: 2503.20104 [cs].

[47] Y. Guo, C. Li, C. Roan, S. Pakhomov, and T. Cohen, "Crossing the "Cookie Theft" Corpus Chasm: Applying What BERT Learns From Outside Data to the ADReSS Challenge Dementia Detection Task," *Frontiers in Computer Science*, vol. 3, Apr. 2021.