Early Detection of Alzheimer's Disease from Speech with Hierarchical Fusion Networks

A Research Proposal

Raftopoulos Michail

National Technical University of Athens, Greece

Abstract

Early detection of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) is crucial for timely intervention and more effective disease management. While automated speech analysis has shown promise for AD detection, current approaches face significant limitations: until recently, most focused on binary AD versus healthy control classification, neglecting the clinically critical MCI stage, models often lack interpretability required for clinical adoption, and generalization across languages remains limited. This research proposes a hierarchical fusion network designed to address these challenges by incorporating multiple interpretable feature modalities—including acoustic, linguistic, paralinguistic, and demographic information—at various levels of abstraction. The hierarchical architecture adds transparency to the model's decision-making process through attention mechanisms while promising high performance on the challenging three-class classification problem (AD/MCI/HC).

1 Introduction

Dementia is a progressive neurodegenerative disorder that causes irreversible brain damage, significantly impairing activities of daily living. It is estimated that it affects over 57.4 million people, a figure that is expected to rise dramatically due to the aging population [13]. Alzheimer's Disease (AD) is the most prevalent form of dementia, accounting for 60-70% of all cases [17].

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. Particularly, 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 [1, 8].

2 Problem Statement

There have been many advancements in the field of AD detection from spontaneous speech. Organizations such as DementiaBank have provided the research community with essential datasets, and many competitions have used the existing datasets to produce valuable benchmarks for researchers to develop and evaluate their models. These efforts have resulted in numerous successful approaches, with reported accuracies often surpassing 90%. [15, 14]. However, several critical limitations prevent clinical translation of these technologies.

Detection of MCI: The majority of previous studies have focused on the binary classification task between AD and Healthy Control (HC) subjects. The inclusion of MCI as a separate class presents a significantly more challenging three-class classification problem, yet it is clinically essential as MCI represents a critical intervention window where treatments may be most effective. [15, 16].

Language Generalization: Currently, most datasets contain only English speech samples. This limits the global applicability of the models, since important markers may differ across languages. [16] showcases that while some languages benefit from multilingual training, others require language-specific models.

Interpretability: Clinical adoption requires model interpretability, as healthcare professionals need to understand which speech characteristics drive diagnostic predictions to integrate findings with other clinical indicators and build professional trust. However, most previous approaches have utilized complex deep learning architectures and/or pre-trained language models, which are often considered black boxes, hindering model interpretability.

These challenges have attracted considerable attention from the research community, leading to a growing number of recent studies addressing the topic.

3 Objectives

Out of the mentioned problems, this project will mainly focus on the detection of MCI and aim for high performance in the according 3-class classification problem (AD/MCI/HC). The model will be developed in a way that preserves an acceptable level of interpretability through the use of various proven interpretable features, fused with a hierarchical architecture. The attention weights at the fusion stage will provide valuable insight on the model's decision making process.

4 Literature Review

4.1 Datasets and Evaluation Challenges

DementiaBank: Most datasets currently available are provided by the DementiaBank database. The largest and most widely used of them is the Pitt corpus, a collection of transcripted audio recordings from various cognitive tasks. 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 transcripted videos of a large variety of cognitive tasks and includes subjects of a wide ethnic and cultural diversity [8].

It should be mentioned that a recent analysis [9] identified the presence of the Clever Hans effect within the Pitt corpus. This phenomenon occurs when models exploit spurious dataset artifacts rather than learning genuine diagnostic patterns, potentially inflating reported accuracies while compromising real-world generalizability. This finding raises important questions about the reliability of some previously reported performance metrics and underscores the need for more robust evaluation protocols.

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 [11]. 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 [12]. These challenges have provided valuable benchmarks and lead to significant advancements in the field.

MultiConAD: The MultiConAD dataset [16] 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 [18] 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 tasts: Sematic 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.

4.2 Relevant Approaches

Table 1 summarizes the most successful approaches for the classification problem in the known 3-class datasets. For the MultiConAD baseline, the selected model was the top-performing one that used only English transcriptions and dense text representations [16]. It should be noted that we could not find any other approaches that train and evaluate on the MultiConAD dataset. In the PROCESS Challenge, the most performant baseline model utilized only eGeMAPS acoustic features and a Random Forest classifier [18]. Some proposed baselines in the paper perform better, but are trained and evaluated on subsets of the dataset. The overall winner of the PROCESS Challenge [4] 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. The best performing submission specific to the classification task [21] achieved an F1 score of 0.696, using a self-developed Digital Linguistic Biomarker (DLB) extractor.

A review of top-performing approaches reveals that much of the literature emphasizes the extraction of informative feature sets as a key factor in achieving high performance. Several notable studies illustrate this trend. For instance, [22] 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, [6] 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. Finally, [7] leverages fine-tuned Whisper models for AD detection in low-resource languages. The authors chose to additionally integrate

speaker background information such as age, gender, and education level, showing that the inclusion of demographic variables leads to significant performance improvements.

Table 1: Most successful approaches and baselines for the 3-class classification problem (AD/MCI/HC).

Study	Dataset	Method	Accuracy	F1 score
[16]	MultiConAD (English only)	SVM + `multilingual-e5-large` (baseline)	0.65	-
[18]	PROCESS	RF + eGeMAPS (baseline)	0.525	0.474
[4]	PROCESS	Whisper + linguistic features + esemble of traditional ML models	-	0.649
[21]	PROCESS	DLB extractor + RF + PCA	-	0.696

4.3 Hierarchical Fusion Networks

Hierarchical architectures have been extensively applied across a wide range of domains, consistently demonstrating strong performance. In particular, they have proven highly effective in affective computing and multimodal learning, where the ability to integrate information at the appropriate levels of granularity—ranging from low-level signals to higher-level abstractions—has been shown to capture complex patterns more accurately [2, 5, 19, 20].

In addition to their proven strong performance, we believe that hierarchical fusion networks can also exhibit a high level of interpretability. If interpretable features are used, and we decide to fuse them by concatenation, one can examine the attention weights at the fusion stage to understand which features contributed the most to the model's decision.

Although hierarchical architectures offer these advantages, our review of the literature revealed no prior approaches that apply them specifically to AD and MCI detection. This absence highlights a promising research gap: by leveraging hierarchical structures, future models could potentially capture speech and language patterns at multiple levels of representation, leading to more accurate and interpretable predictions.

5 Methodology

5.1 Dataset

Since the main focus of this project is the detection of MCI, the PROCESS dataset is considered to be the most suitable, as it provides the most modern and well-structured dataset for this task. However, the data does not seem to be publicly available yet, so a special request from the organizers will be needed.

Another viable option is the MultiConAD dataset. Its large collection of samples can allow larger models to be trained, and its multilingual nature can facilitate experiments with joint multilingual training. The majority of the data can be acquired though a request to the DementiaBank organization.

5.2 Architecture

The proposed model will include audio and language modalities, with additional feature sets explored to identify the most effective combination. Since this approach might add some layers of complexity, the development process will be divided into 2 stages, with the first one focusing on a simpler model to act as a strong baseline, and the second on expanding that baseline with additional features.

5.2.1 Stage 1: Model Baseline

The baseline model will utilize audio and linguistic features, as these are the most significant for the problem at hand. For acoustic features, the eGeMAPS feature set [3] was selected, due to its interpretability and proven effectiveness in paralinguistic and clinical speech analysis. These features will be extracted using the OpenSMILE toolkit. For the linguistic modality, the raw audio will be transcribed using a pre-trained ASR model (e.g., Whisper), and the resulting text will be encoded using a model like BERT to generate embeddings. Although in most DementiaBank datasets manual transcripts are provided, they will be omitted as they would not be available in a real-world clinical setting.

The selected modalities will be aligned at the word level, to provide fine-grained temporal information. Then, they will be fused through concatenation, followed by processing via a Transformer Encoder. Those representations will then be aggregated to the recording level

using a pooling layer and subsequently fed into the classification head. Lastly, the model's hyperparameters will be optimized either manually or algorithmically, depending on the available computational resources.

5.2.2 Stage 2: Additional Feature Integration

The second stage focuses on experimentation with additional feature sets, inspired by the literature review. There is also interest in experimenting with different abstraction levels for the already existing modalities, such as phrase-level or recording-level embeddings. The potential candidates can be summarized as follows:

- Algorithmically extracted paralinguistic features.
- Psycholinguistic features and affective information.
- Demographic data.
- Phrase-level fusion of embeddings.
- Recording-level fusion of embeddings.

Through an ablation study, we will identify the optimal combination of feature sets and assess the relative importance of each set in contributing to model performance. After the optimal feature combination is identified, the model's hyperparameters may be re-optimized to ensure peak performance.

5.3 Possible Extensions

In addition to the primary objectives of the project, some potential extensions could be explored, depending on the availability of time and resources. Firstly, given the focus on the interpretable traits of the proposed architecture, an in-depth interpretability assessment would enhance the robustness of this study. Secondly, if the MultiConAD dataset is used, there would be interest in training the model in multilingual data and evaluating its performance on low-resource languages, such as Greek.

5.4 Expected Outcomes

This project is expected to demonstrate that hierarchical fusion networks can effectively address the three-class classification problem of AD, MCI, and healthy controls while preserving interpretability. By systematically integrating and evaluating diverse feature sets, the study aims to identify which modalities and levels of representation contribute most to robust performance. We hope that this approach will result in a model that is not only highly accurate but also transparent, aligning with the critical needs of the healthcare community.

References

[1] "2018 Alzheimer's Disease Facts and Figures". In: *Alzheimer's & Dementia* 14.3 (Mar. 2018), pp. 367–429.

[2] Georgios Chatzichristodoulou et al. *MEDUSA: A Multimodal Deep Fusion Multi-Stage Training Framework for Speech Emotion Recognition in Naturalistic Conditions*. Sept. 2025. arXiv: 2506.09556 [cs].

[3] Florian Eyben et al. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". In: *IEEE Transactions on Affective Computing* 7.2 (Apr. 2016), pp. 190–202.

[4] Yifan Gao, Long Guo, and Hong Liu. *Leveraging Multimodal Methods and Spontaneous Speech for Alzheimer's Disease Identification*. Feb. 2025. arXiv: 2412.09928 [cs].

[5] Efthymios Georgiou, Charilaos Papaioannou, and Alexandros Potamianos. "Deep Hierarchical Fusion with Application in Sentiment Analysis". In: *Interspeech 2019*. ISCA, Sept. 2019, pp. 1646–1650.

[6] Fasih Haider et al. *Affective Speech for Alzheimer's Dementia Recognition*. May 2020.

[7] Kaichen Jia et al. "Whisper-Based Multilingual Alzheimer's Disease Detection and Improvements for Low-Resource Language". In: ().

[8] Alyssa M. Lanzi et al. "DementiaBank: Theoretical Rationale, Protocol, and Illustrative Analyses". In: *American Journal of Speech-Language Pathology* 32.2 (Mar. 2023), pp. 426–438.

[9] Yin-Long Liu et al. *Clever Hans Effect Found in Automatic Detection of Alzheimer's Disease through Speech*. June 2024. arXiv: 2406.07410 [eess].

[10] Saturnino Luz and Fasih Haider. "Detecting Cognitive Decline Using Speech Only: The ADReSSO Challenge". In: ().

[11] Saturnino Luz et al. *Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge*. Aug. 2020. arXiv: 2004.06833 [eess].

[12] Saturnino Luz et al. "An Overview of the ADReSS-M Signal Processing Grand Challenge on Multilingual Alzheimer's Dementia Recognition Through Spontaneous Speech". In: *IEEE Open Journal of Signal Processing* PP (Mar. 2024), pp. 1–12.

[13] Emma Nichols 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". In: *The Lancet Public Health* 7.2 (Feb. 2022), e105–e125.

[14] Benjamin S. Runde, Ajit Alapati, and Nicolas G. Bazan. "The Optimization of a Natural Language Processing Approach for the Automatic Detection of Alzheimer's Disease Using GPT Embeddings". In: *Brain Sciences* 14.3 (Feb. 2024), p. 211.

[15] Arezo Shakeri and Mina Farmanbar. "Natural Language Processing in Alzheimer's Disease Research: Systematic Review of Methods, Data, and Efficacy". In: *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring* 17.1 (Feb. 2025), e70082.

[16] Arezo Shakeri, Mina Farmanbar, and Krisztian Balog. *MultiConAD: A Unified Multilingual Conversational Dataset for Early Alzheimer's Detection*. Feb. 2025. arXiv: 2502.19208 [cs].

[17] Marcos Vinícius Ferreira Silva et al. "Alzheimer's Disease: Risk Factors and Potentially Protective Measures". In: *Journal of Biomedical Science* 26 (May 2019), p. 33.

[18] Fuxiang Tao et al. *Early Dementia Detection Using Multiple Spontaneous Speech Prompts: The PROCESS Challenge*. Dec. 2024. arXiv: 2412.15230 [cs].

[19] Ilias Triantafyllopoulos, Georgios Paraskevopoulos, and Alexandros Potamianos. *Depression Detection in Social Media Posts Using Affective and Social Norm Features*. Mar. 2023. arXiv: 2303.14279 [cs].

[20] D. Xezonaki et al. *Affective Conditioning on Hierarchical Networks Applied to Depression Detection from Transcribed Clinical Interviews*. https://arxiv.org/abs/2006.08336v1. June 2020.

[21] Shibingfeng Zhang et al. "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.

[22] Maryam Zolnoori, Ali Zolnour, and Maxim Topaz. "ADscreen: A Speech Processing-based Screening System for Automatic Identification of Patients with Alzheimer's Disease and Related Dementia". In: *Artificial intelligence in medicine* 143 (Sept. 2023), p. 102624.