MULTIMODAL ALZHEIMER'S DETECTION: SPEECH & TEXT

DIEGO TORIBIO, JAEHYEON PARK, LAMIAH KHAN, WONGEE HONG, & NOLAN GRIFFITH | ADVISED BY SAM KEENE | VIP-381-D: BIOENGINEERING

THE COOPER UNION

INTRODUCTION

Alzheimer's dementia (AD) is the world's leading neuro-degenerative disease, affecting roughly 55 million people and progressing silently for years before clinical diagnosis. Subtle changes in everyday conversation—rhythm, pitch, word choice—often appear long before costly imaging or invasive tests confirm decline, making speech a promising low-cost screen. Leveraging the balanced doctor–patient dialogues of ADReSSo-2021 [1], we ask whether a single model that both listens to prosody and reads transcribed language can flag AD more reliably than audio-only or text-only approaches.

OBJECTIVE

Using the **ADReSSo-2021** corpus, we extend last semester's audio- and text-only models to a single **multimodal** approach. Recordings are encoded with frame-level eGeMAPS speech features and transformer-based embeddings from the generated transcripts. Fusing these vectors, we retrain our classifiers to test whether joining *how* words are spoken with *what* they say improves Alzheimer's-dementia detection beyond either modality alone.

FRAMEWORK

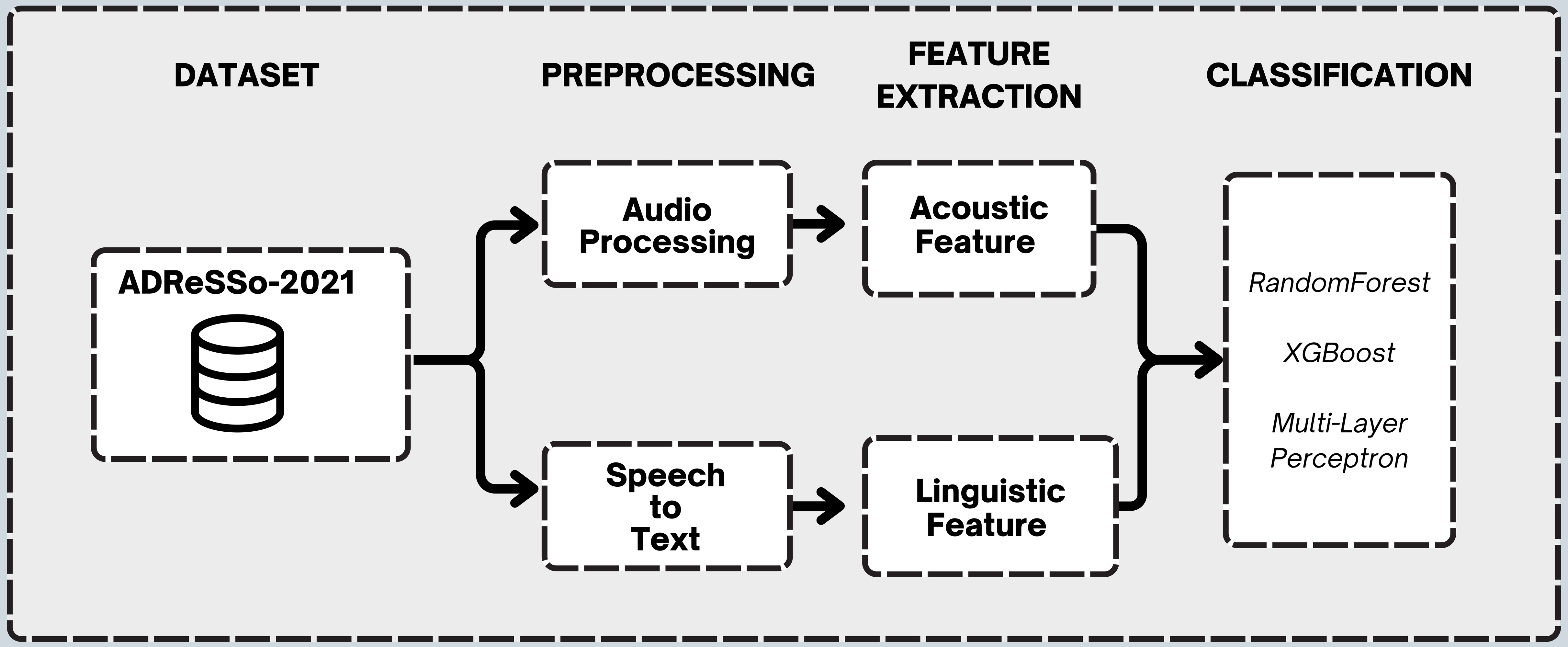

Figure 1. High-level pipeline for multimodal Alzheimer's detection: ADReSSo-2021 audio is split into an audio branch(paralinguistic processing → acoustic features) and a text branch (speech-to-text → linguistic features). The resulting vectors are fed to tree-based or neural classifiers for AD vs CN prediction.

METHODS

Each recording is first resampled to 16 kHz with *librosa* [3] and cropped to **patient-only speech** using the provided speaker time-stamps. We then branch into two complementary streams:

- **Audio stream** (*how* **the words are spoken)**
  - **openSMILE** extracts eGeMAPS paralinguistic features, while a pretrained wav2vec 2.0 model provides a higher-level acoustic embedding [4–6].
- **Text stream** (*what* **is said)**
  - Whisper-large v2 generates a transcript, and DistilBERT converts each sentence to a compact language embedding [7, 8].

The resulting 1024-D audio vector and 768-D text vector are z-scored and concatenated into a 1 792-D feature that gives the classifier both vocal delivery and lexical content in one shot.

Feature Set	Tool / Model	Dimension
eGeMAPS (prosody)	openSMILE	88
wav2vec 2.0 (speech)	Base model	1024
DistilBERT (language)	CLS-token mean	768
Fusion (concat)	-	1880

Table 1. Dimensionality of each feature stream—eGeMAPS (88 D, openSMILE), wav2vec 2.0 (1 024 D), DistilBERT (768 D)—and their 1 880-D concatenated vector.

FEATURE EXTRACTION

Audio path: Patient speech is windowed at 100 ms and 250 ms with 0 % or 50 % overlap. Every frame yields 25 eGeMAPS descriptors; mean ± std pooling forms an 88-D prosodic vector. The same frames feed wav2vec 2.0, whose hidden states are averaged to a 1 024-D embedding.

Text path: Whisper produces time-stamped transcripts; sentence boundaries guide DistilBERT, and the sentence embeddings are averaged to a single 768-D semantic vector.

Fusion: The eGeMAPS, wav2vec, and DistilBERT vectors are normalized, concatenated (1 880 features), and supplied to the classifier, giving it both the rhythm of speech and the meaning of words.

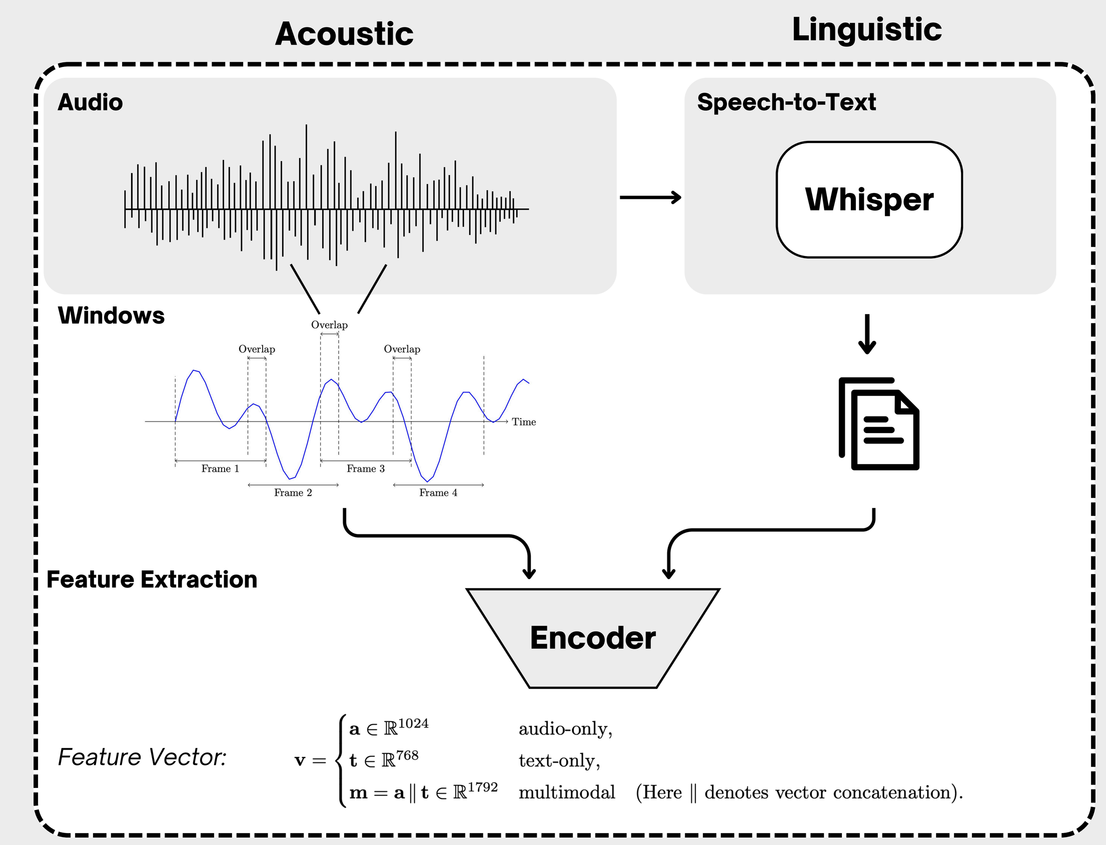

Figure 2. Feature-extraction pipeline: audio is windowed, embedded, and joined with Whisper-derived text embeddings to form the final feature vector v.

RESULTS

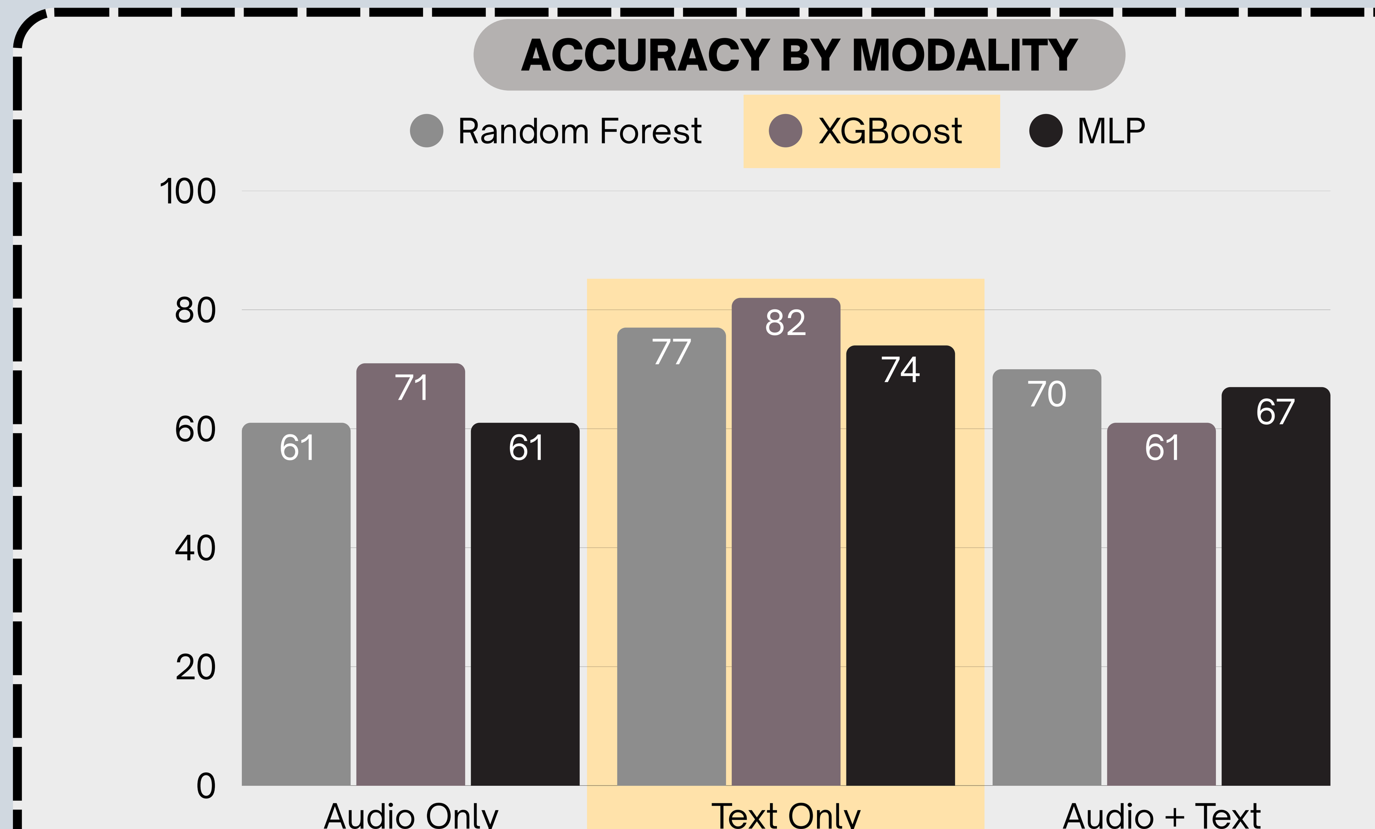

Figure 3 ranks the best classifier from each stream. **Text-only (DistilBERT + XGBoost)** tops the chart at **82 % accuracy / 0.83 F1**, confirming that word-level information is the single strongest cue. Adding speech boosts the audio baseline: the **multimodal fusion model** reaches ***70 % accuracy / 0.74 F1***, edging out the **audio-only** pipeline (***67 % / 0.70 F1***) built on wav2vec 2.0 features. Precision and recall trail the corresponding accuracies by < 3 pp in every case, so each model's wins and losses are evenly distributed between AD and CN speakers.

Figure 3. Best accuracy achieved per modality—audio only, text only, and fused audio + text—broken down by classifier.

CONCLUSION

Automatic transcripts carry the clearest Alzheimer's signal in this study; paralinguistic cues alone lag behind. A simple late-fusion of audio and text lifts the audio baseline but still sits below the text-only ceiling, hinting that smarter integration—joint attention layers or larger speech encoders—may be needed to unlock the full value of acoustic patterns. Even so, the fusion results show that speech features can add robustness without hurting accuracy, pointing toward richer multimodal designs as the next step for conversational dementia screening. For source code and replication details, scan the QR code in the lower-right corner.

REFERENCES

[1] https://dementia.talkbank.org/ADReSSo-2021/
[2] McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, and Nieto O, "librosa: Audio and music signal analysis in python," in Proceedings of the 14th python in science conference, vol. 8. Citeseer, 2015, pp. 18–25. [Google Scholar]
[3] Eyben F, Wöllmer M, and Schuller B, "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462. [Google Scholar]
[4] Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, and Quatieri TF, "A review of depression and suicide risk assessment using speech analysis," Speech Communication, vol. 71, pp. 10–49, 2015. [Google Scholar]
[5] Syed ZS, Sidorov K, and Marshall D, "Automated screening for bipolar disorder from audio/visual modalities," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 39–45. [Google Scholar]
[6] "Dementia," Sep 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia
[7] Tosto G, Gasparini M, Lenzi G, and Bruno G, "Prosodic impairment in alzheimer's disease: assessment and clinical relevance," The Journal of neuropsychiatry and clinical neurosciences, vol. 23, no. 2, pp. E21–E23, 2011. [PubMed] [Google Scholar]
[8] Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, Devillers LY, Epps J, Laukka P, Narayanan SS et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," IEEE transactions on affective computing, vol. 7, no. 2, pp. 190–202, 2015. [Google Scholar]
[9] Luz S, Haider F, Fuente S, Fromm D, MacWhinney B, "Detecting cognitive decline using speech only: The ADReSSo Challenge"
[10] Hui J, "Speech Recognition — Feature Extraction MFCC & PLP"

openSMILE audio feature extraction