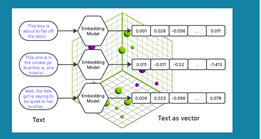
A Novel AI-powered Pipeline for Alzheimer's Disease Classification using Spontaneous Speech and Vector Embeddings



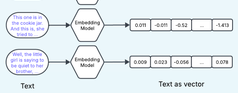
ABSTRACT

Alzheimer's disease (AD) is a growing global public health concern, making **early detection** crucial for effective intervention. **Spontaneous speech** offers a promising, low-cost, and non-invasive biomarker for AD diagnosis. In this study, we use Large Language Models (LLMs) APIs to **automate transcriptions** of spontaneous speech audio files, and generate **vector embeddings to capture the semantic meanings** of the transcription files. We then trained machine learning classification models to diagnose AD patients from healthy controls, and regression models to predict the MMSE cognitive scores. We found the **speaker diarization** significantly improved data accuracy and model performance. Manual transcription yielded the best classification results, while AssemblyAI and Whisper (with GPT-4o post-processing) also performed well. By automating audio transcription, generating vector embeddings, and applying machine learning models, we aim to provide a **scalable and accessible AI-driven solution** for early AD detection.

INTRODUCTION

Early diagnosis of Alzheimer's is crucial for a timely intervention and better care planning. Traditional diagnosis include Cognitive and Behavior Tests, Brain Imaging, Biomarkers & Lab Tests, etc.
Limitations: need access to medical facilities, invasive and expensive tests, risk of missing early-stage symptoms

Spontaneous speech, captured in natural real-life settings, holds potential as a digital biomarker for accessible and non-invasive AD diagnosis and screening. Both acoustic and linguist features has been extensively studied but heavily depend on domain knowledge and hand-crafted transformation. There is a huge potential in using AI and especially the large language model in understanding the latent semantic meanings in the speech data.



Large Language Models are trained on vast datasets to understand the semantic meanings behind words, sentences or images. Text can be turned into an **embedding, a vector (list) of floating point numbers** in high dimension space. The distance between two vectors measures their relatedness: Small distances suggest high relatedness and large distances suggest low relatedness.

RESEARCH OBJECTIVES

- Design an AI-powered pipeline to automatically process and transcribe spontaneous speech recordings
- Use vector embedding to capture semantic features of transcriptions
- Train machine learning models using acoustic features from audio files and semantic features from embeddings
- Compare the classification models and regression models performances

DATA

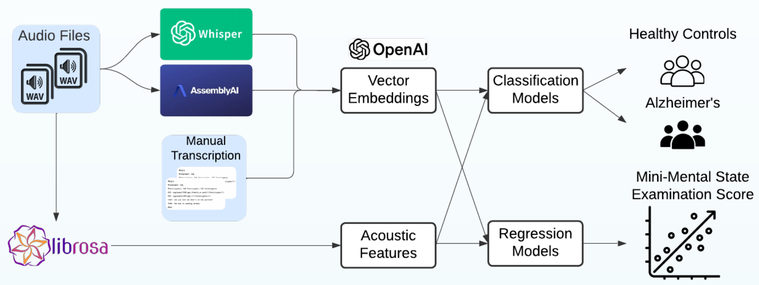
- Audio recordings of the "**Cookie Theft**" picture description, part of the DementiaBank's Pitt Corpus.
- Data pre-processed acoustically and mitigated common biases.
- Task #1: **AD classification**
- Task #2: **MMSE score prediction**

Age	Train Data						Test Data					
	AD			non-AD			AD			non-AD		
	M	F	MMSE (sd)	M	F	MMSE (sd)	M	F	MMSE (sd)	M	F	MMSE (sd)
[50, 55]	1	0	30.0 (n/a)	1	0	29.0 (n/a)	1	0	23.0 (n/a)	1	0	28.0 (n/a)
[55, 60]	5	4	16.3 (4.9)	5	4	29.0 (1.3)	2	2	18.7 (1.0)	2	2	28.5 (1.2)
[60, 65]	3	6	18.3 (6.1)	3	6	29.3 (1.3)	1	3	14.7 (3.7)	3	3	28.7 (0.9)
[65, 70]	6	10	16.9 (5.8)	6	10	29.1 (0.9)	3	4	23.2 (4.0)	3	4	29.4 (0.7)
[70, 75]	6	8	15.8 (4.5)	6	8	29.1 (0.8)	3	3	17.3 (6.9)	3	3	28.0 (2.4)
[75, 80]	3	2	17.2 (5.4)	3	2	28.8 (0.4)	1	1	21.5 (6.3)	1	1	30.0 (0.0)
Total	**24**	**30**	**17.0 (5.5)**	**24**	**30**	**29.1 (1.0)**	**11**	**13**	**19.5 (5.3)**	**11**	**13**	**28.8 (1.5)**

Table 1: ADReSSo training and testing datasets. M = male, F = female, n/a = not applicable.

https://dementia.talkbank.org/ADReSS-2020/

METHODS

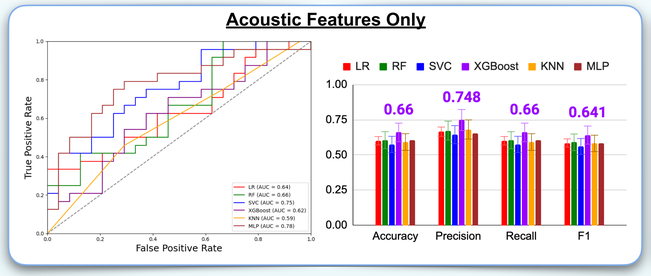


- **Acoustic Features Extraction** with Librosa for 41 features, including Pitch (F0) statistics (4), Mel-frequency cepstral coefficients(MFCCs) (26), Spectral features(2), Spectral Rolloff(2), Spectral Bandwidth(2), Zero Crossing Rate(2), Root Mean Square Energy(2), Speech rate estimation(1)
- **Semantic Features Extraction** with OpenAI Vector Embedding models:
  - text-embedding-3-small (1536 dimensions) and text-embedding-3-large (3072 dimensions)
  - **Audio Transcriptions**:
    - Manual transcription: filter out "PAR" user
    - By AssemblyAI API: filter with utterance speaker
    - By Whisper API: raw or post process with prompt
- **Classification Models** from scikit-learn (see below)
- **Regression Models** from scikit-learn (see below)

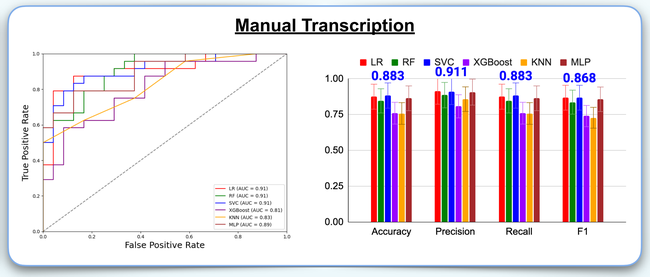
RESULTS

Classification Models from scikit-learn: Logistic Regression (LR), Random ForestClassifier (RF), Support Vector Classifier(SVC), XGBClassifier(XGBoost), KNeighborsClassifier(kNN), MLPClassifier(MLP); With text-embedding-3-small (small embedding) by default, unless marked otherwise;

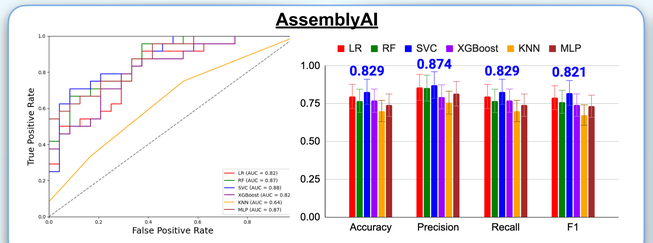
Acoustic Features Only



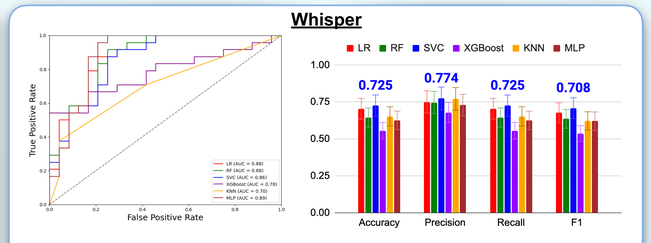
Manual Transcription



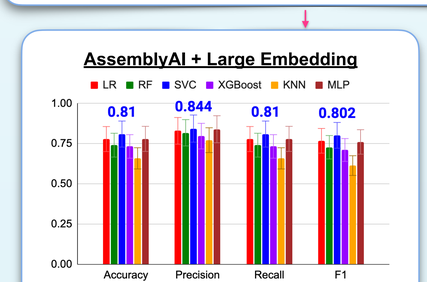
AssemblyAI



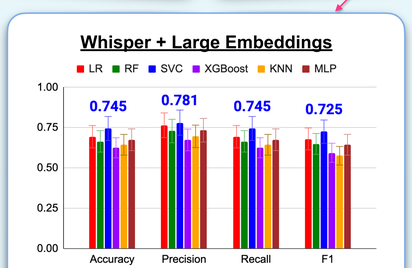
Whisper



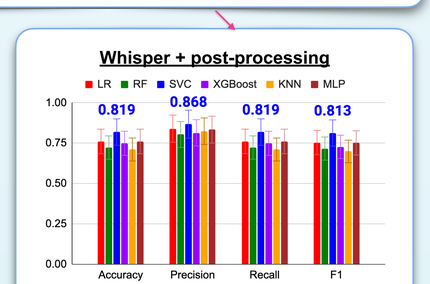
AssemblyAI + Large Embedding



Whisper + Large Embeddings



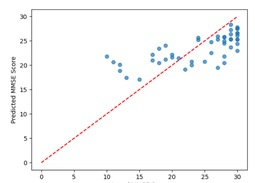
Whisper + post-processing



Regression models for MMSE value prediction: Ridge, RandomForestRegressor, XGBRegressor

Transcript Type	Model	RMSE (R^2)
Manual Transcript	Ridge	5.23 (0.26)
	RandomForestRegressor	4.98 (0.33)
	XGBRegressor	4.83 (0.37)
AssemblyAI Transcript	Ridge	5.34 (0.22)
	RandomForestRegressor	**4.51 (0.45)**
	XGBRegressor	4.70 (0.40)
Whisper Transcript with Post Processing	Ridge	6.14 (-0.03)
	RandomForestRegressor	5.57 (0.16)
	XGBRegressor	5.59 (0.15)

Table 1: Regression model performance with different transcription types; all using text-embedding-3-small embedding.



MMSE Prediction with AssemblyAI transcript and small embedding

Discussions

Speaker Diarization — Important!
- Separates speakers in audio files, crucial to distinguishing participant speech from interviewer speech in the The ADReSSo Challenge dataset.
- **Pyannote.audio** (PyTorch-based) was tested but performed poorly on local laptop; may explorer pyannoteAI service next

Transcription & Speaker Identification
- **Manual transcription**: Labels speakers as "INV" (Investigator) and "PAR" (Participant), leading to the **best classification results**
- **AssemblyAI:** Provides speaker-labeled utterances, resulting in strong classification performance
- **Whisper:** Does **not** separate speakers by default, but applying a **GPT-4o post-processing LLM prompt** achieves results comparable to AssemblyAI
- Whisper Turbo model generated better transcription than the Whisper Tiny model

Acoustic vs Semantic Features:
- Acoustic Features: Extracted 41 features with Librosa library
- **Performance Comparison**: Models using semantic features with vector embeddings outperform those using acoustic features.
- **Feature Combination**: Merging acoustic and semantic features did not enhance performance due to differences in feature group sizes.

Classification and Regression Models Performance
- Models are trained by 10-fold cross-validation and evaluated on test set
- **Support Vector Classifier (SVC)** consistently outperforms other classification models
- **RandomeForestRegressor** based on AssemblyAI transcript outperforms other regression models

Embedding Size Considerations
- Due to the relatively small audio file sizes, **text-embeddings-3-large** (a higher-cost model) does not significantly improve performance

Challenges:
- Experimenting with different API-based models and/or commercial AI services to balance accuracy, speed, and cost
- Optimizing engineering decisions for machine learning model parameter tuning to enhance model performances

CONCLUSION & FUTURE WORK

We have successfully developed an AI-based pipeline for Alzheimer's classification and MMSE score prediction, demonstrating the Large Language Model (LLM) based vector embedding is a viable approach for AD detection and assessment, using recorded spontaneous speech.

Looking ahead, I am excited to expand this research with the following ideas:
- Test the pipeline on datasets with varied age and gender distributions.
- Explore embedding models from different LLMs.
- Apply the pipeline on non-English spontaneous speech recordings.
- Evaluate the pipeline's effectiveness for other speech-sensitive conditions such as aphasia, depression, PTSD, and Parkinson's disease
- Deploy the machine learning model to web site and evaluate with senior users
- Contribute audio recordings to DementiaBank to strength their collection, as requested by the DementiaBank researchers

REFERENCES

- Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of Neurology, 51(6), 585-594.
- NIA AG03705 and AG05133
- Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2020). Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge https://arxiv.org/abs/2004.06833
- Agbavor, F., & Liang, H. (2022). Predicting dementia from spontaneous speech using large language models. PLOS Digital Health, 1 (12), e0000168. doi: 10.1371/journal.pdig.0000168