Detect Alzheimer from Speech

Intern: Zhiyuan Chen (Jennifer)
Mentor: Vijay Ravi

- Introduction to the problem
- Database
- Baseline Experimental Details
- Baseline results
- Previous work
- What we’re currently doing

ADReSSo Challenge

Alzheimer's Dementia Recognition through Spontaneous Speech (only)

2020 ADReSS Challenge:

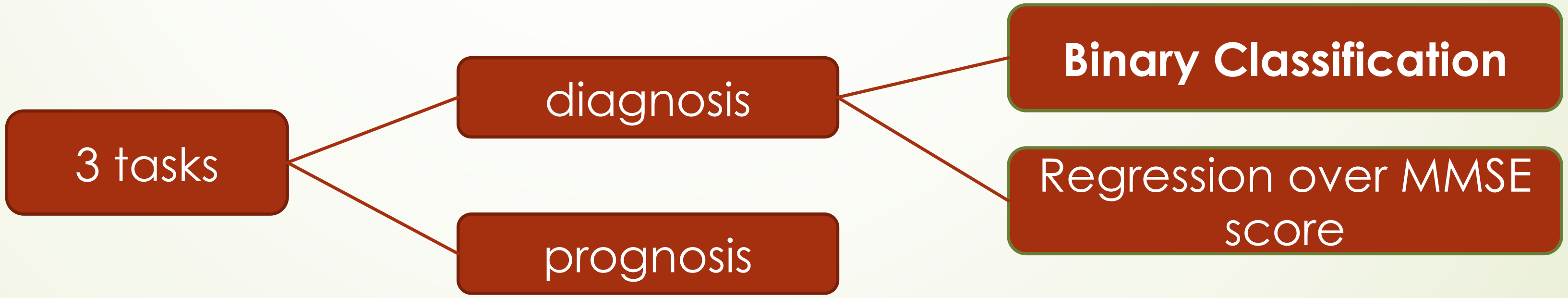
- speech recording + manual script

  acoustic feature + linguistic feature

2021 ADReSSo Challenge:

- speech recording only

  focuses on the acoustic characteristic of speech



2021 database

Speech recordings of 2 distinct datasets:

- (prognosis) AD patients performing semantic fluency task for baseline visit
- (diagnosis) **picture descriptions by AD patients and control group**

Speech sample: 237 participants

Train		Test
ad	cn	71
87	79	

166 training + 71 testing (7/3 split)

Training data: .wav files + segmentation timestamps

Diagnosis Baseline

Features

Acoustic: apply sliding window with 100ms to obtain frames => eGeMAPS features => ADR method

Linguistic: Google Cloud-based Speech Recogniser => convert transcripts into CHAT format => MOR function => EVAL and FREQ commands

5 Methods

Method	Decision Tree	KNN	Linear Discriminant Analysis (LDA)	Random Forest	SVM
Parameter	Leaf size: {1 to 20}	K: {1 to 20}		50 trees Leaf size: {1 to 20}	Linear kernel Box constraints: {0.1 to 1.0}

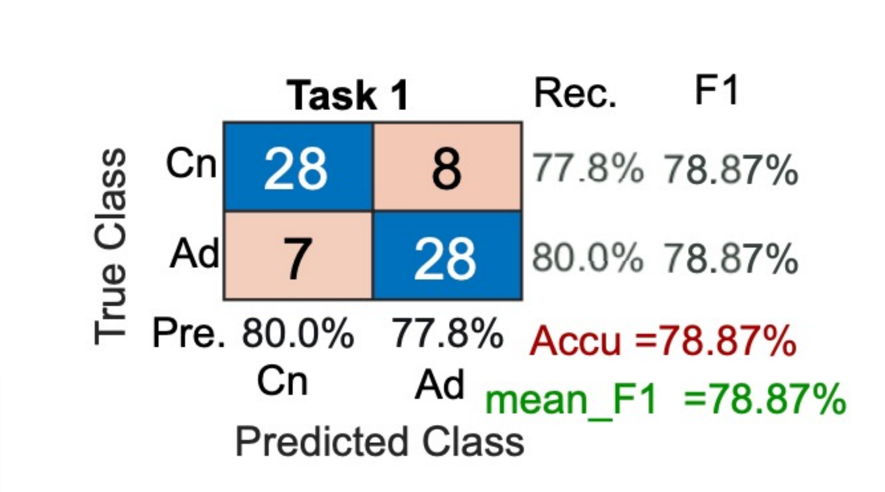
Baseline Results

- Best performing classifier
  - Cross-validation: DT (78.92% acoustic, 72.89% linguistic)
  - Test set: SVM (77.46% linguistic)

	Classifier	LDA	DT	SVM	RF	KNN	mean (sd)
CV	Acoustic	62.65	**78.92**	69.28	65.06	65.06	68.19 (6.4)
	Linguistic	72.29	**72.89**	72.89	75.90	65.06	71.81 (4.0)
Test	Acoustic	50.70	60.56	**64.79**	63.38	53.52	58.59 (6.2)
	Linguistic	76.06	74.65	**77.46**	73.24	59.15	72.11 (7.4)

- Late fusion of the acoustic and linguistic models improves the accuracy on the test set

  => Baseline accuracy of 78.87%



Previous Work

- 2020 best performance paper

• 1st: 89.6% accuracy, focuses solely on linguistic feature.

Utilizes domain knowledge, process the transcript based on pause and the use of "uh", then feed into BERT

• 2nd: 85.42% accuracy, acoustic + linguistic

Different BERT alternatives for every word of every transcript

2021 challenge

Paper1: outperforms the baseline model by 2.82%

- proposed a new method

(combination of engineered features + audio representations)

- pipeline

  - 3 approaches for feature extraction

  - 4 models: LR, SVM, NN, DT (default value for all model parameters)

- Feature extraction approaches

  - approach1: conventional acoustic features

    Verbal: fundamental frequency, jitter, shimmer

    Non-verbal: MFCC

  - approach2: pre-trained acoustic embeddings (deep neural models)

    wav2vec 2.0 (self-supervised audio representation model)

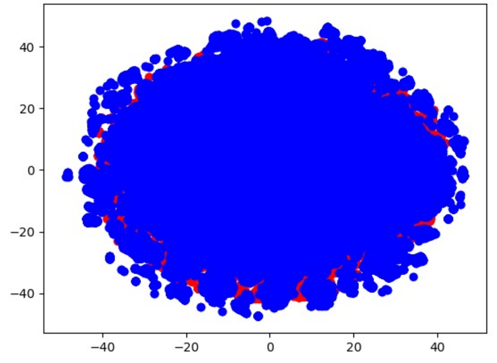
  - approach3: best method

    combination of state-of-the-art self-supervised techniques + domain knowledge

What we’re doing

- Current steps:

1. Segment the audio files, then extract MFCC features
2. t-SNE visualization for MFCC



3. SVM model on MFCC feature

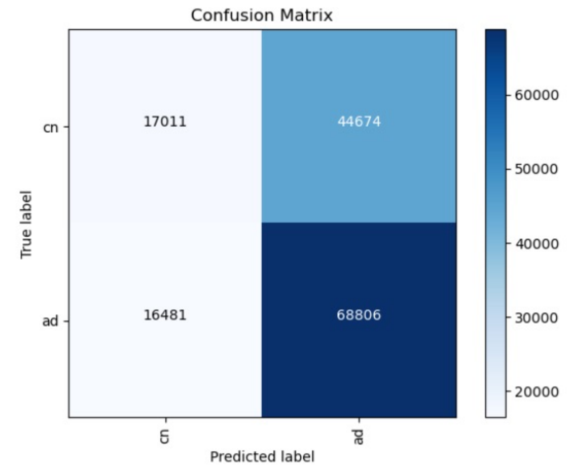
   - provide metrics for 3 different trials on training set

   - predict on frame level, then use majority voting to have the results of segment level and speaker level

Trial 1 result

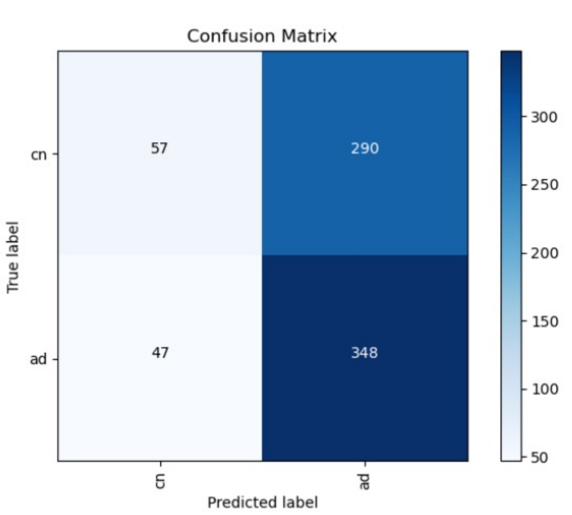
Frame level

```
Accuracy score for frame_SVM: 0.583900
Recall score for frame_SVM: 0.806758
Precision score for frame_SVM: 0.606327
F-1 score for frame_SVM: 0.692328
```



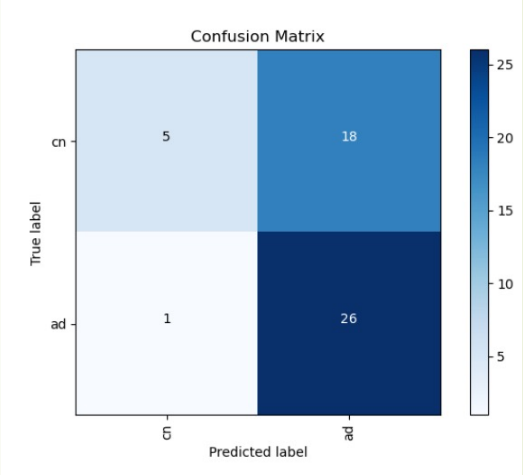
Segment level

```
Accuracy score for segment_SVM: 0.545822
Recall score for segment_SVM: 0.881013
Precision score for segment_SVM: 0.545455
F-1 score for segment_SVM: 0.673766
```



Speaker level

```
Accuracy score for speaker_SVM: 0.620000
Recall score for speaker_SVM: 0.962963
Precision score for speaker_SVM: 0.590909
F-1 score for speaker_SVM: 0.732394
```



- Next steps:

  - produce SVM probability logits, average across 3 trials

  - do the same for eGeMAPS, ComParE16, VQUAL, entropy

Thank you!