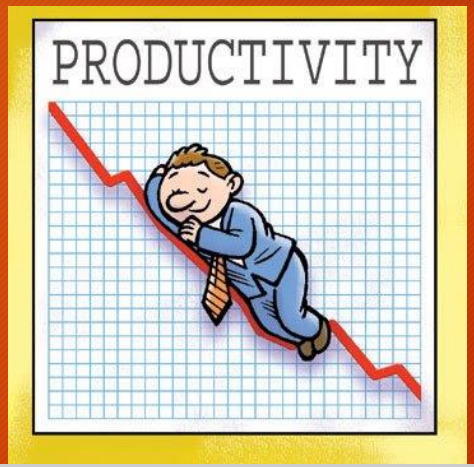
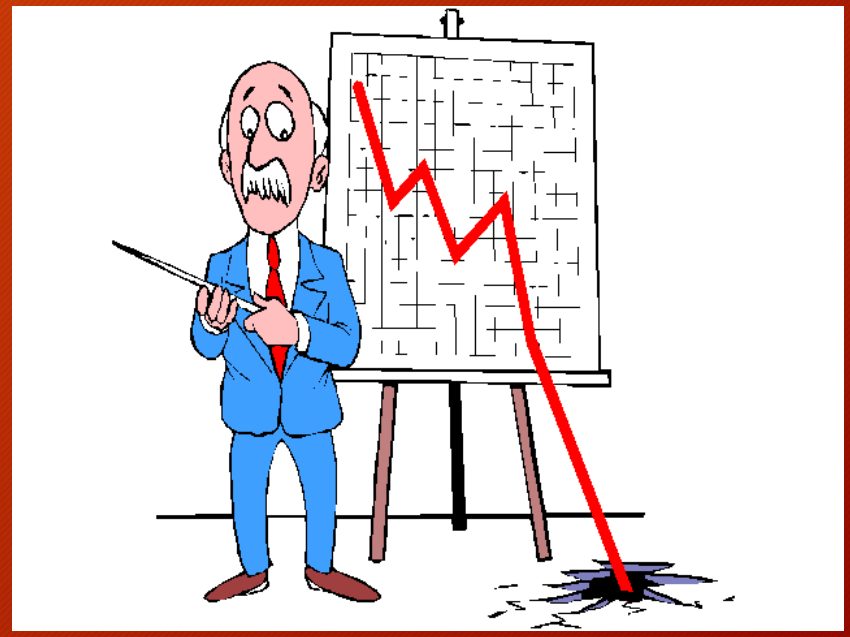
Depression Diagnosis from Speech

Chan Jun Wei
A0112084U
Advisor: Prof Ooi Wei Tsang

Problem



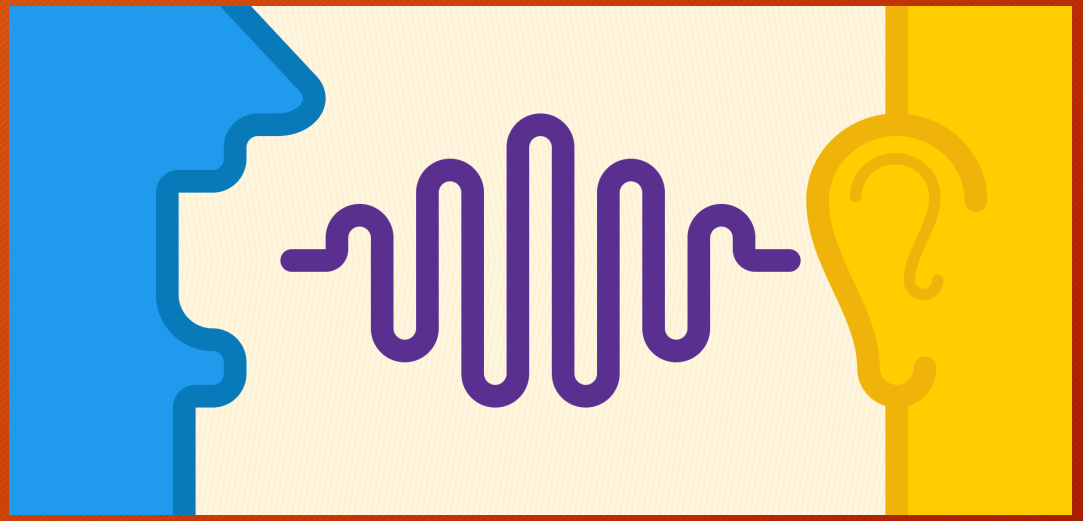


Problem

- Depression is one of the most **under-diagnosed** illnesses even though it is one of the most treatable mental illnesses.
- Reason:
  - Unwilling to seek help
  - Hide their feelings
  - Lack of objective measure

Idea

- Diagnose depression using speech
- “Diminished, prosodic and monotonous speech always has a strong co-relationship with depression.”
- We hope to accurately detect depression from speech.



Audio/Visual Emotion Challenge and Workshop (AVEC) 2016

- Depression Classification Sub-challenge (DCC)
  - Dataset:
    - Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ)
  - Baseline:
    - Classification: Mean F1 of 0.5
    - Regression: RMSE of 6.74
  - Best:
    - Classification: Mean F1 of 0.81
    - Regression: RMSE of 5.31
- Can we achieve a mean F1 of 0.8 on AVEC2016 dev set?

PHQ-8	Depression Level
0 ~ 4	Minimal Depression
5 ~ 9	Mild Depression
10 ~ 14	Moderate Depression
15 ~ 19	Moderately Severe Depression
20 ~ 24	Severe Depression

Related Studies

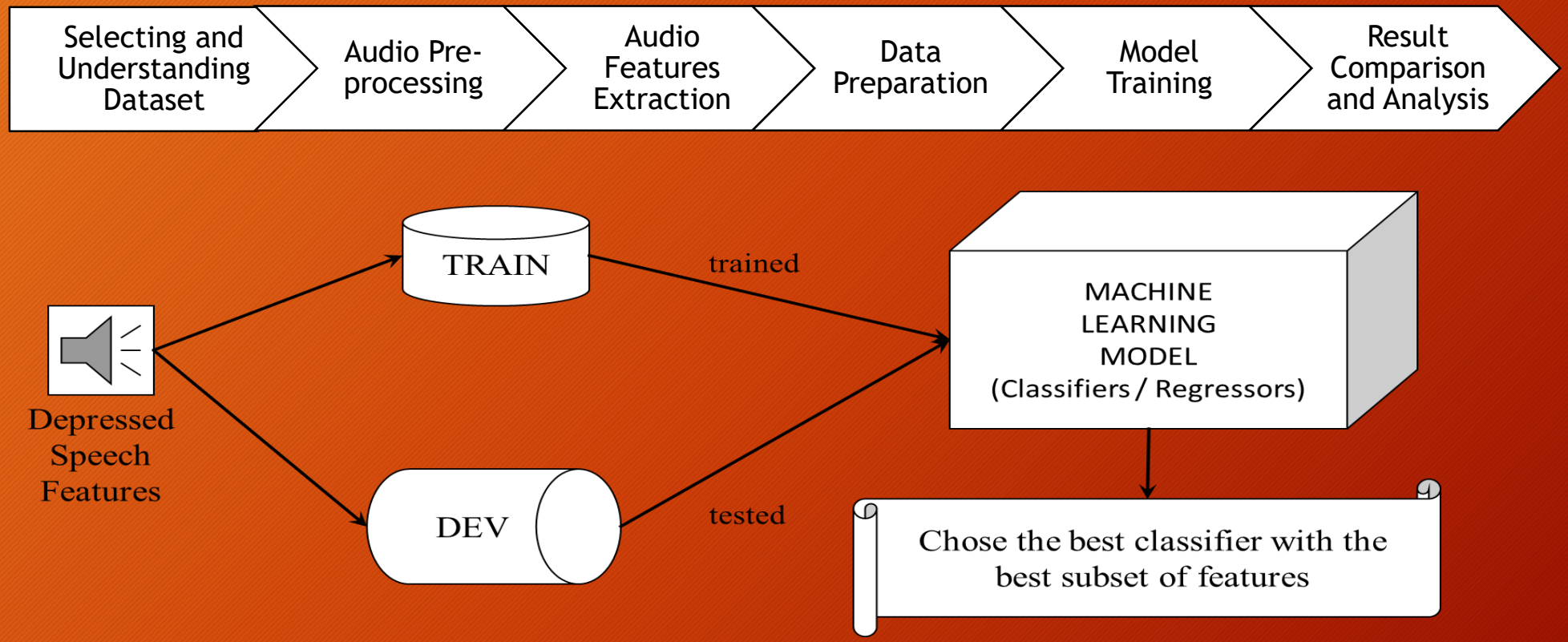
- AVEC 2016
  - Baseline
  - DepAudioNet
  - MIT Lincoln Laboratory (MITLL)
  - Pampouchidou's Study
  - SCUBA
- Other Studies

Overview of Work Done

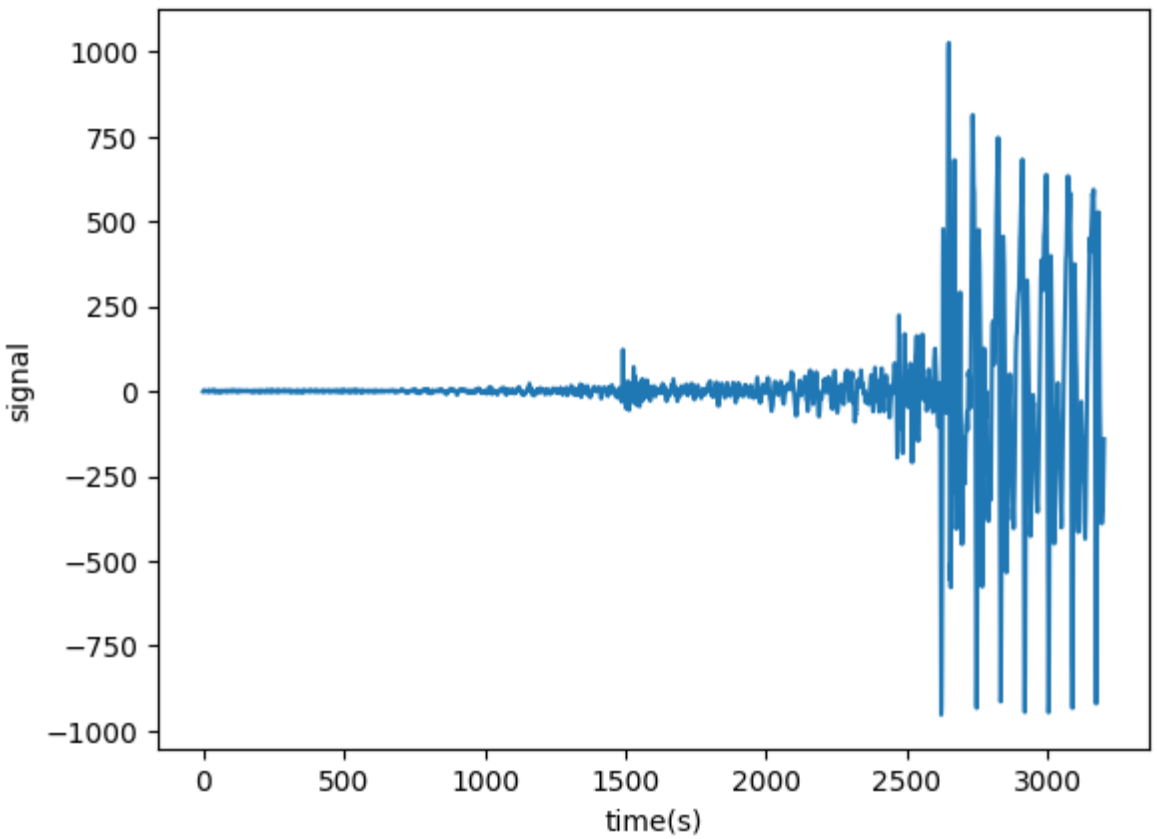
- Binary classification Task (Depressed or not)
- Regression Task (PHQ-8 score)
- Multi-class classification Task (PHQ-8 Depression Level)
- The need of Normalization
- The need of Feature Selection

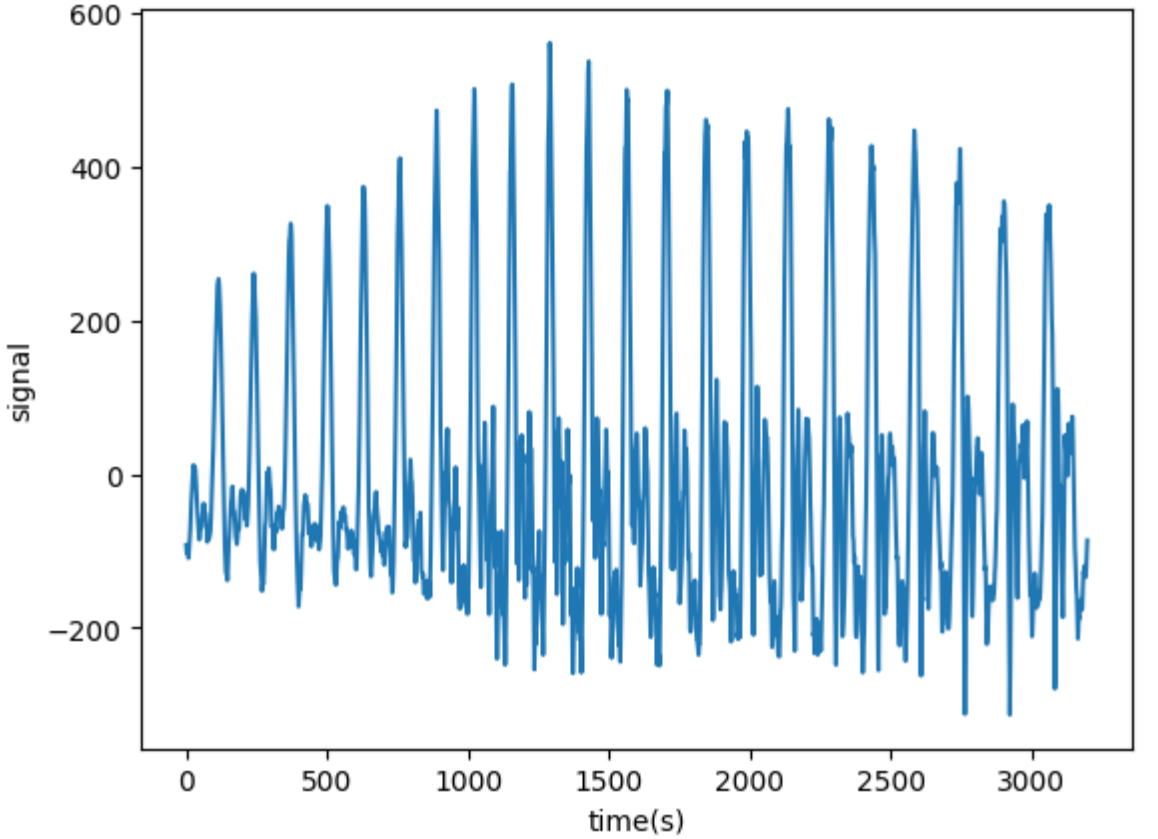
Method

- Process



Understanding Dataset

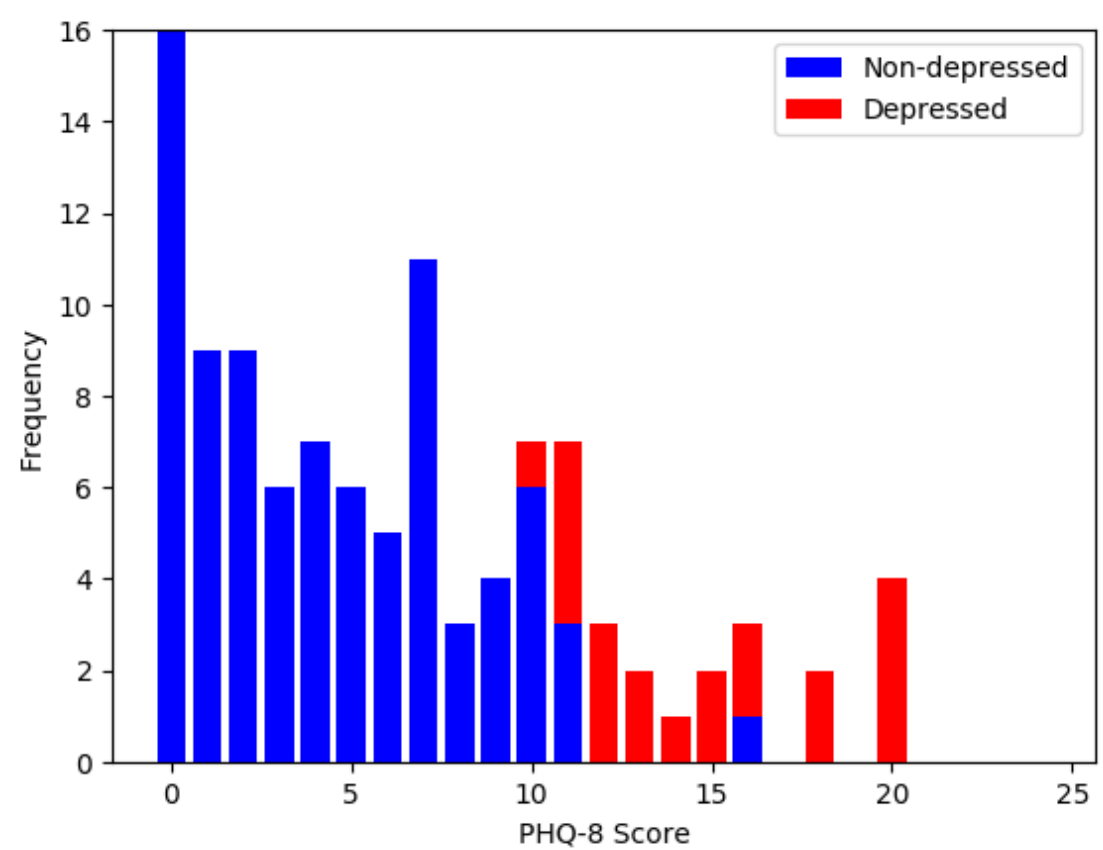




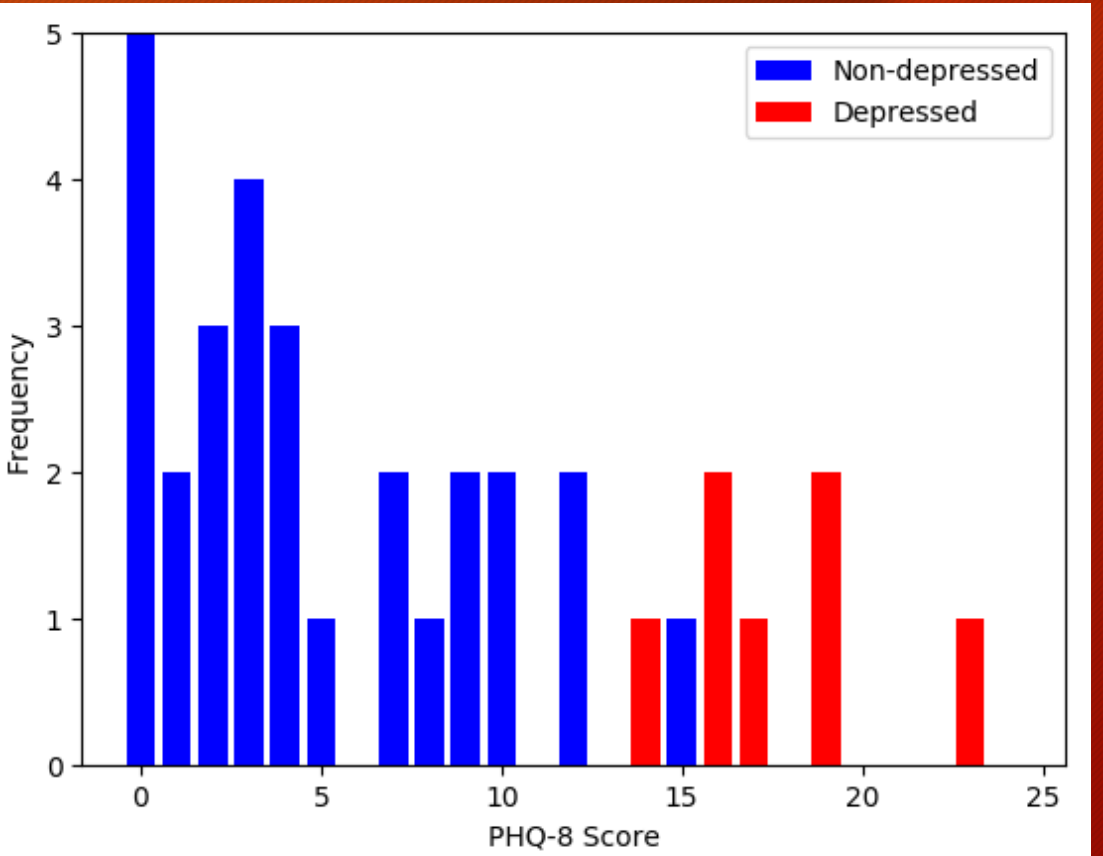
Non-depressed Sample

Depressed Sample

Understanding Dataset

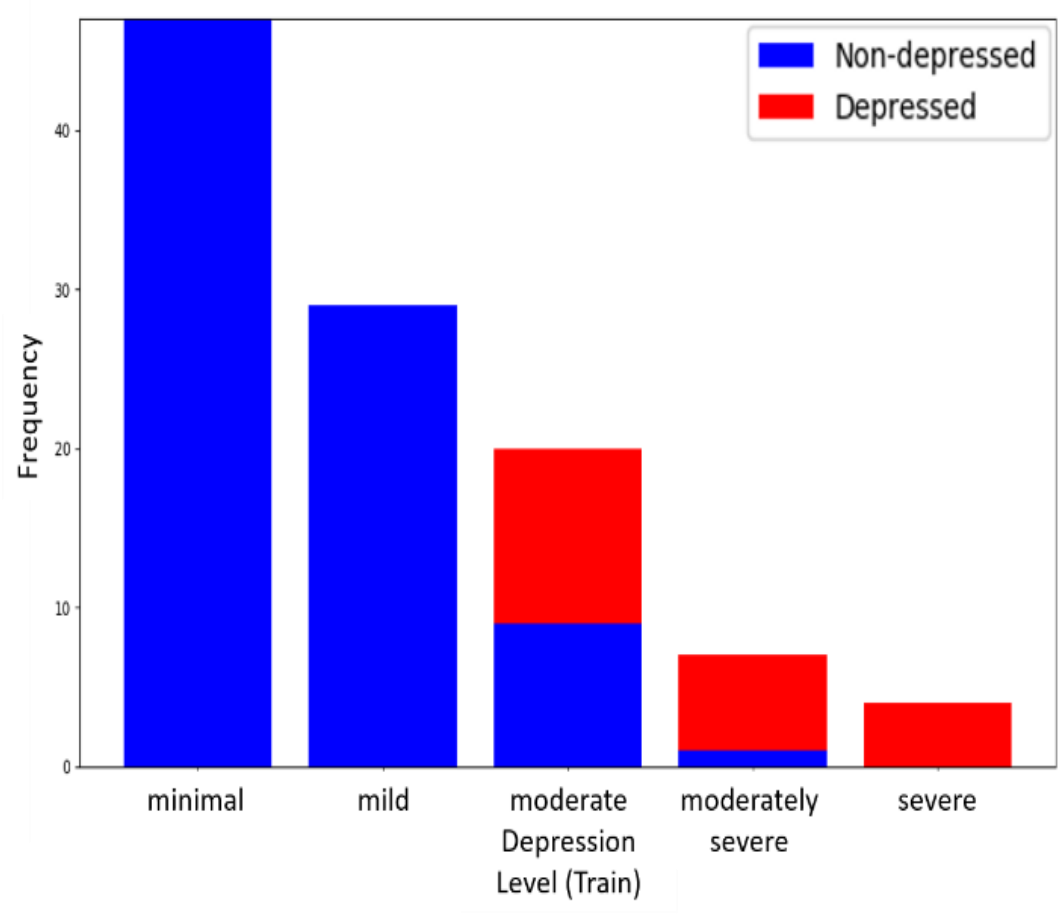


Train Set

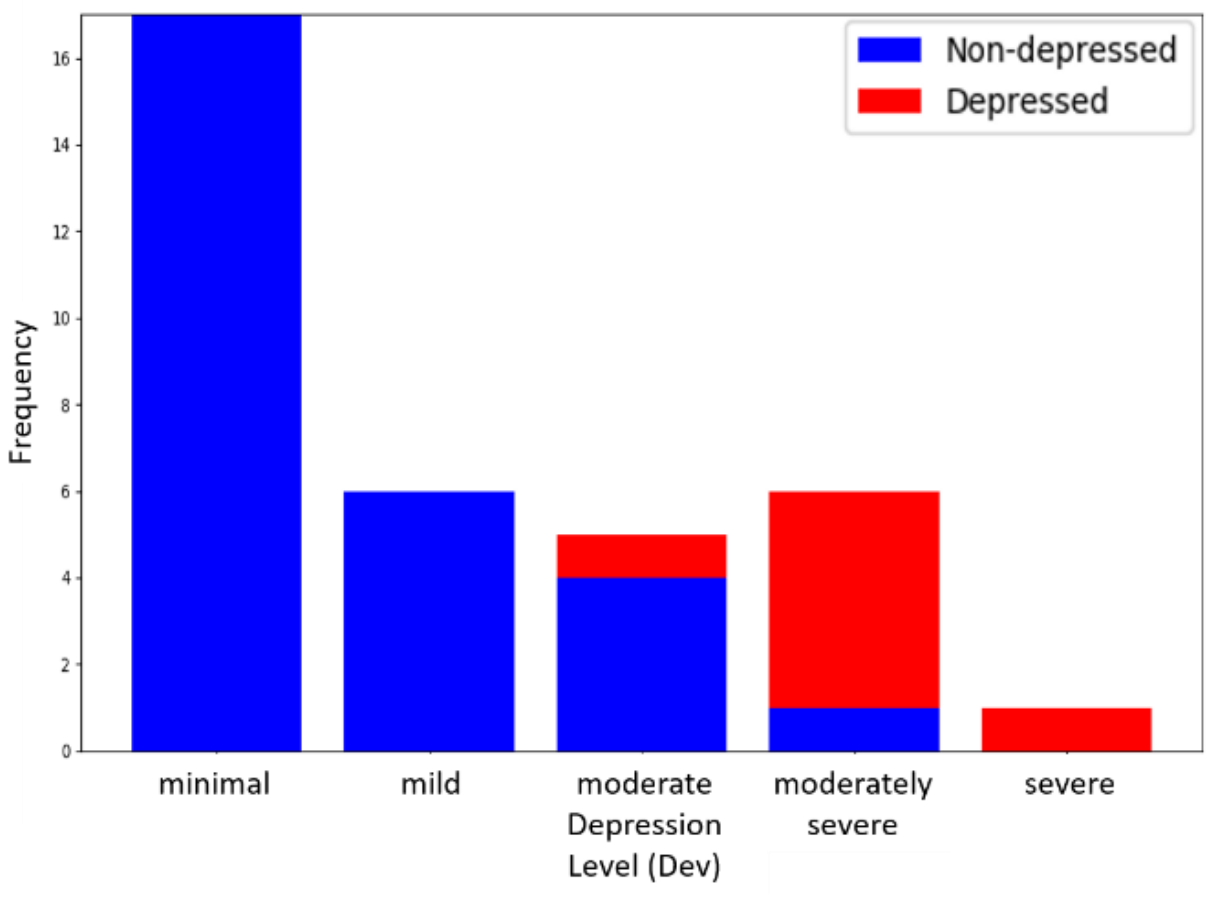


Dev Set

Understanding Dataset



Train Set



Dev Set

Audio Pre-Processing

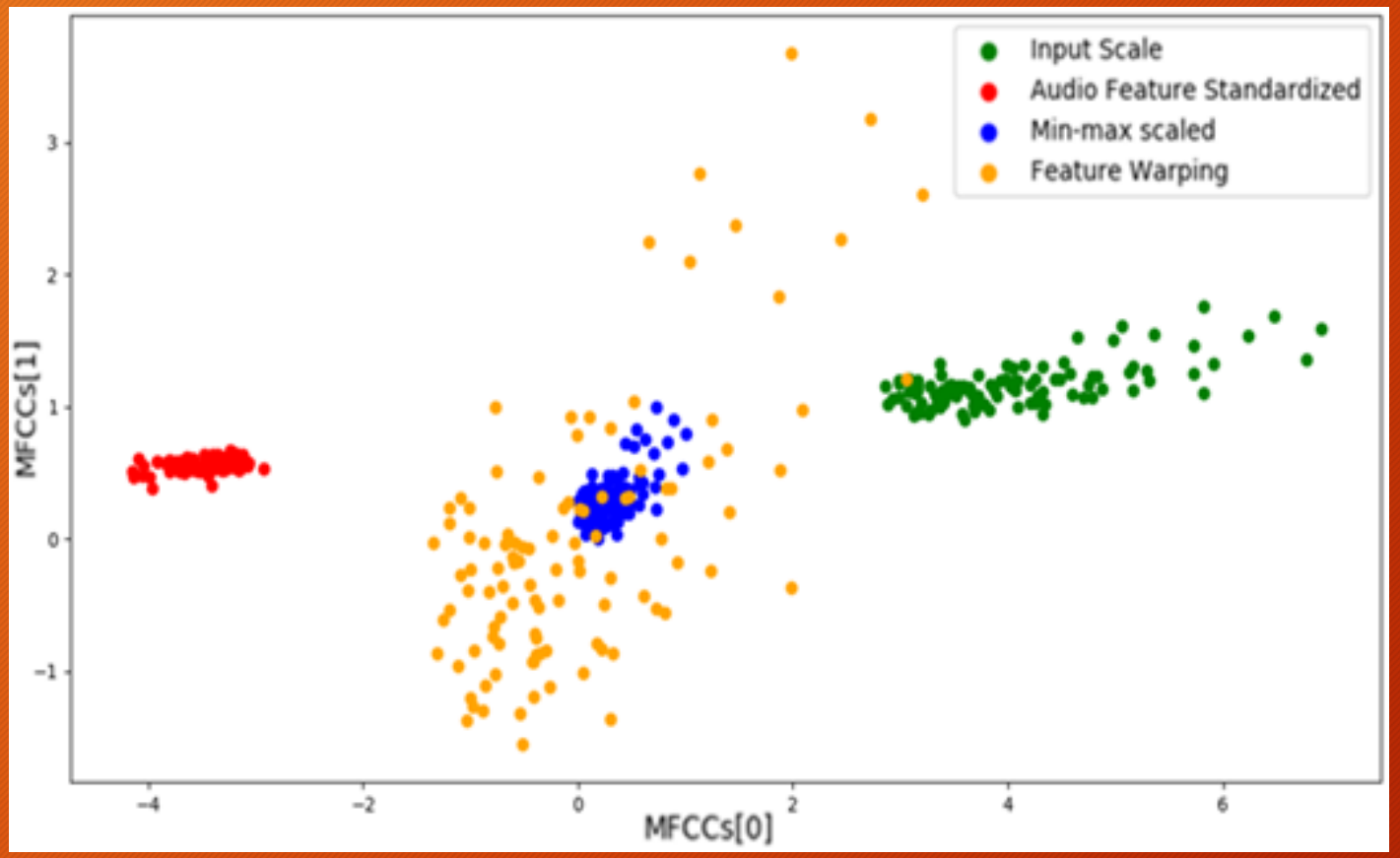
- Obtain original raw audio.
- Noise Reduction using FFMPEG and SOX.
- Obtain Speech Segments using transcript.
- Merge them into one audio.

Start time	Stop time	Speaker	Value
60.0	61.3	Ellie	How are you doing today?
62.3	63.2	Participant	Good.

Audio Features Extraction

Index	Audio Feature Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies, it can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9~21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22~33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The std of the 12 Chroma coefficients.

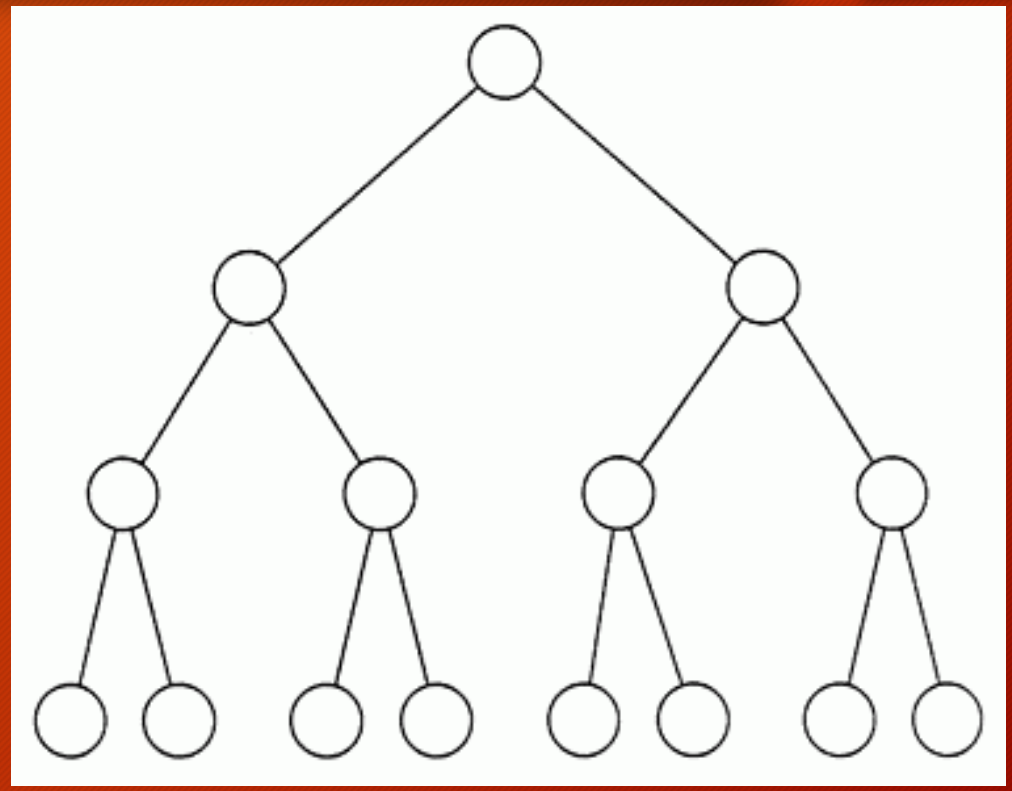
Data Preparation (Normalization)



- We propose: Audio Feature Standardization

Data Preparation (Feature Selection)

- We propose Audio Feature Selection via Complete Search
- We examined Relief, Fisher, CIFE, CMIM, MRMR, MIFS, ICAP, FCBF, CFS and GINI.

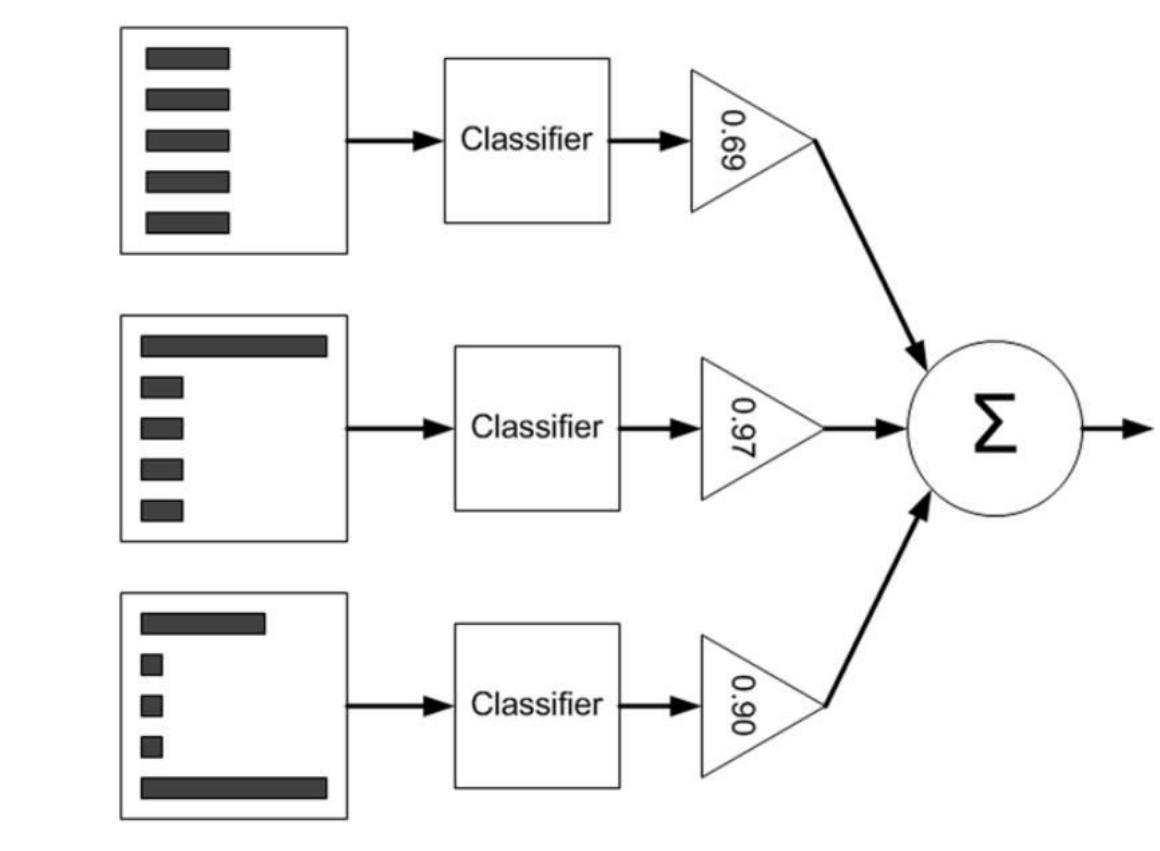


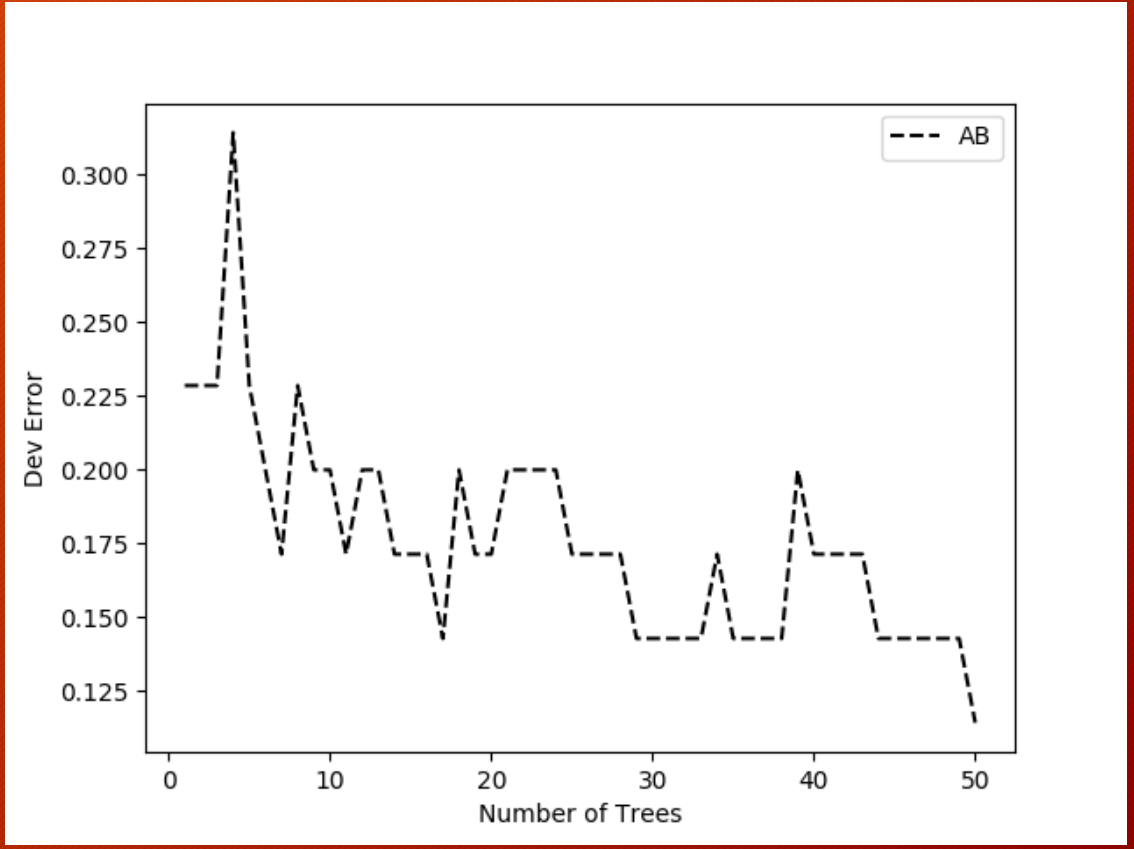
Model Training

Names	Description
Naïve Bayes (NB)	NB estimates based on the assumption that all features are independent with each other.
Gaussian Process (GP)	GP focuses on stochastic processes which generalize the probability distribution for functions.
Support Vector Machine (SVM)	SVM tries to maximize the linear margin between the two separating hyperplane and thus focus on creating the largest possible distance between the two hyperplanes.
K-Nearest Neighbours (KNN)	KNN is constructed based on the assumption of all the similar data will generally exist closer to each other.
Decision Tree (DT)	DT contains a decision tree which performs the classification task by sorting them based on feature values.
Random Forest (RF)	RF is an ensemble of decision trees trained with different subsets of the training examples.
AdaBoost (AB)	AB is an ensemble of learners with different hypothesis by making changes to training examples. It maintains a set of weights over the learners.

Model Training

- AdaBoost (AB)





Result (Normalization)

- Effects of Audio Feature Standardization

Audio Features	F1 Score (Without Normalization)	Classifier	F1 (With Audio Feature Standardization)	Classifier
Zero-Crossing Rate	0 (0.89)	GP-DP	0 (0.89)	GP-DP
Energy	0 (0.89)	GP-DP	0 (0.89)	GP-DP
Entropy of Energy	0 (0.89)	GP-DP	**0.36 (0.88)**	AB
Spectral Centroid	0.18 (0.85)	AB	0 (0.89)	GP-DP
Spectral Spread	0.18 (0.85)	AB	0.14 (0.78)	AB
Spectral Entropy	0 (0.89)	GP-DP	**0.25 (0.78)**	AB
Spectral Flux	0.33 (0.86)	AB	0.25 (0.90)	KNN
Spectral Rolloff	0 (0.89)	GP-DP	0 (0.89)	GP-DP
MFCCs	0.2 (0.87)	AB	**0.36 (0.88)**	AB
Chroma Vector	0.2 (0.87)	AB	**0.25 (0.90)**	GP-DP
Chroma Deviation	0 (0.89)	GP-DP	**0.13 (0.76)**	AB

Result (Binary Classification)

Audio Features	Model	Mean F1	F1
Mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation	AB	0.82	0.71 (0.93)
Normalized Mean of Spectral Centroid, Spectral Entropy, Spectral Rolloff, MFCCs, Chroma Vector, Chroma Deviation	AB	0.77	0.6 (0.93)
Mean and Std of Spectral Centroid, Spectral Spread, Spectral Flux	AB	0.80	0.67 (0.93)
Normalized Mean and Std of Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Flux	AB	0.70	0.53 (0.87)

Result (Audio Classification Comparison)

Related Study	Model	Audio Features	Normalization	Optimization	Mean F1
Proposed Method	**AB**	**Mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation**	**None**	**None**	**0.82**
Baseline	Linear SVM with SGD	F0, VUV, NAQ, QOQ, H1H2, PSP, MDQ, peak-Slope, Rd, MCEP0-24, HMPDM0-24, HMPDD0-12	None	Grid Search for hyper-parameter optimization	0.50
DepAudioNet	CNN and LSTM	Mel-scale filter bank feature	Batch Normalization	Random Sampling	0.61
MITLL	Gaussian Staircase Model	Correlation structure of formant tracks, Correlation structure of dMFCCs, Lower Vocal Tract (VT) Resonance Pattern, peak-to-rms	Min-max Scaling	Z-scoring, PCA	0.57
Pampouchidou	Decision Fusion Model which is implemented using DT	Baseline, Rd conf, Formants 1-3, the deltas & delta-deltas for F0 & MFCCs, Pause Ratio, Voiced Segment Ratio, Speaking Ratio, Mean Laughter Duration, Mean Delay to answer question, Mean Duration of Pauses, Max Duration of Pauses & Fraction of overall pauses	Min-max Scaling	Feature Selection accesses based on the improvement of modal by removing features.	0.73
SCUBA	G-PLDA	Ivector (MFCC)	None	MIM	0.73

Result (Binary Classification Comparison)

Related Study	Modality	F1	Mean F1
Proposed Method (AB)	Audio	0.71 (0.93)	0.82
Baseline	Audio	0.41 (0.58)	0.50
Baseline	Audio-Video	0.58 (0.86)	0.72
DepAudioNet	Audio	0.52 (0.70)	0.61
MITLL	Audio	-	0.57
MITLL	Audio-Video-Semantic	-	0.81
Pampouchidou	Audio-Gender	0.59 (0.87)	0.73
Pampouchidou	Audio-Video-Text-Gender	0.62 (0.91)	0.77
SCUBA	Audio	0.57 (0.89)	0.73
SCUBA	Audio-Video	0.63 (0.89)	0.76

Result (Regression Comparison)

Related Study	Modality	RMSE	MAE
Proposed Method (AB)	Audio	6.43	5.32
Baseline	Audio	6.7418	5.3566
Baseline	Audio-Video	6.6212	5.5222
MITLL	Audio	6.38	5.32
MITLL	Audio-Video-Semantic	5.31	4.18
SCUBA	Audio	6.7334	5.8237

Result (Summary of All Tasks)

Tasks	Baseline	Best (Audio)	Best (All)	Proposed Result
Binary Classification	Mean F1 of 0.5	Mean F1 of 0.73	Mean F1 of 0.81	Mean F1 of 0.82
Regression	RMSE of 6.74	RMSE of 6.38	RMSE of 5.31	RMSE of 6.43
Multi-class Classification	None	None	None	Mean F1 of 1

Contribution

- AB trained by Mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation provides mean F1 of 0.82 and RMSE of 6.43.
- In multi-class classification, AB gives mean F1 of 1.
- The optimal Audio Feature Set is chosen using Audio Feature Selection via Complete Search.
- Audio Feature Standardization should be used with care.


Q&A
Thank You!

Hidden (MFCC Algorithm)