modality.ai

# Towards Remote Differential Diagnosis of Mental and Neurological Disorders using Automatically Extracted Speech and Facial Features

Vanessa Richter[1], Michael Neumann[1] and Vikram Ramanarayanan[1,2]
[1]Modality.AI, Inc., San Francisco, CA, USA
[2]University of California, San Francisco, San Francisco, CA, USA
vikram.ramanarayanan@modality.ai

## Introduction

- The development of **clinically valid, automatically extracted digital biomarkers** for neurological & mental disorders has the potential to (a) **aid clinicians** in achieving quicker and more reliable diagnoses, and (b) offer **fast** and **objective insights** into patients' states.
- **Research Questions:**
  - How **accurately** can a **machine learning classifier differentially distinguish** between multiple disorders: **depression (DEP), schizophrenia (SCHIZ), bulbar symptomatic (BS)** and **bulbar presymptomatic (BP) ALS**?
  - Which **modalities** and **features** are **most useful** for this multi-class classification task?

## Data & Methods

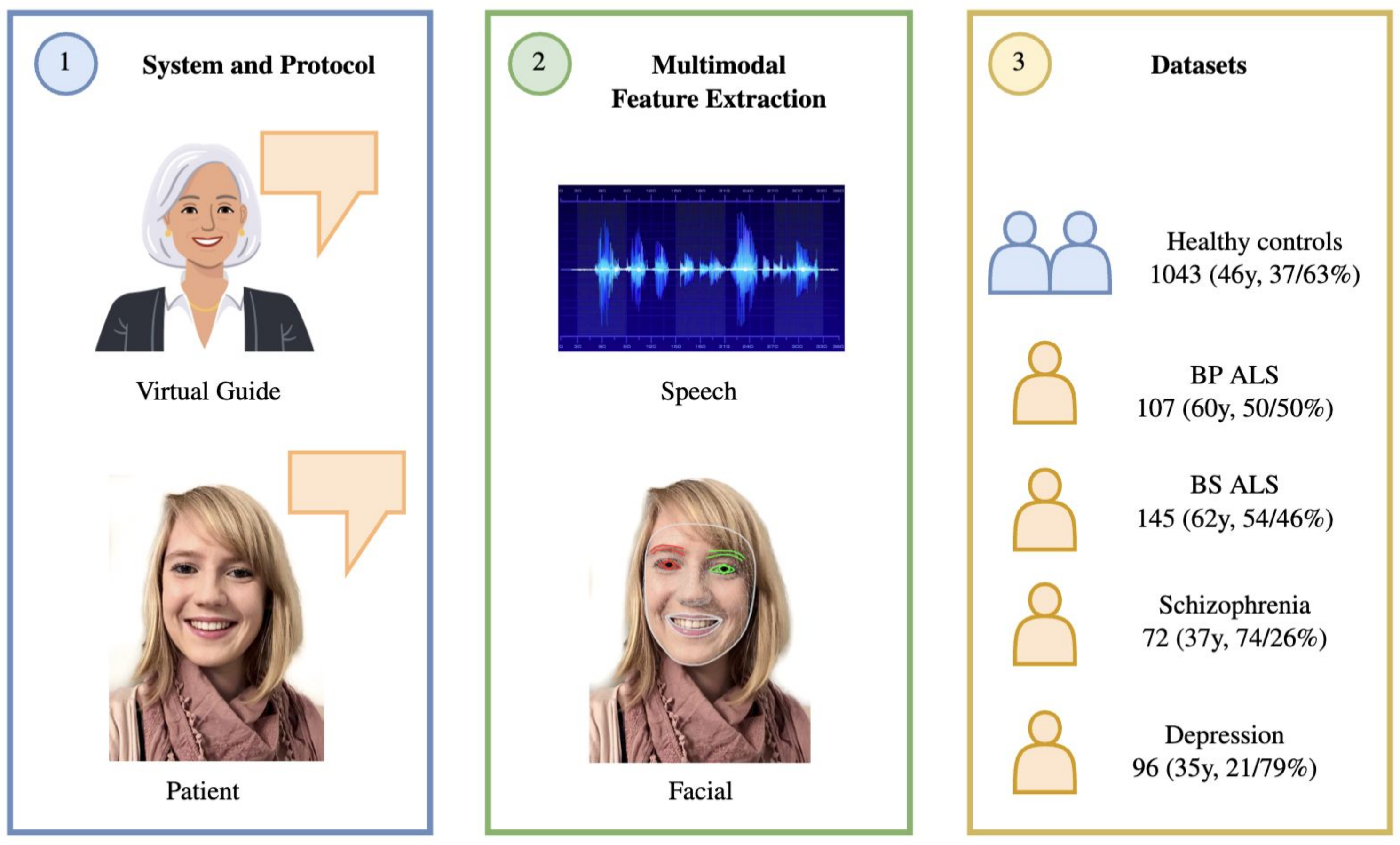


**Figure 1.** Data collection and datasets: The number of sessions is displayed in (3) below the cohorts; mean age in years (y) and male/female ratio in brackets.

| # | Cluster domain | Metrics | Tasks |
|---|---|---|---|
| 1 | Energy | SNR | all |
| 2 | Timing alignment | CTA | all |
| 3 | Timing, pauses | PPT | all |
| 4 | Timing, speaking (1) | articulation/speaking duration | Picture Description |
| 5 | Energy & articulation skills | SNR, syl.rate, syl.count & cTV | DDK |
| 6 | Timing, speaking (2) | articulation/speaking rate/time | SIT_{5,9} |
| 7 | Timing, speaking (3) | articulation/speaking rate/time | SIT_{7,11,13,15}, Bamboo task |
| 8 | Voice quality (DDK skills) | HNR, jitter & shimmer | DDK |
| 9 | Voice quality (periodicity) | HNR | all except DDK |
| 10 | Voice quality (amplitude variation) | shimmer | all except DDK |
| 11 | Voice quality (frequency variation) | jitter | all except DDK |
| 12 | Frequency (mean, min) | min & mean F0 | all |
| 13 | Frequency (max, std) | max & std F0 | all |

**Table 1(a) & (b).** Audiovisual features and clusters of task-metric combinations derived using hierarchical clustering.

| # | Cluster domain | Metrics | Tasks |
|---|---|---|---|
| 1 | Lip movement (1) | speed, acc. & jerk measures | all except DDK |
| 2 | Lip width | mean & max lip width | all |
| 3 | Mouth opening | mean & max lip aperture, mouth surface area | all |
| 4 | Lip movement (2) | speed, acc. & jerk metrics | DDK |
| 5 | Jaw movement (1) | speed, acc. & jerk metrics | DDK |
| 6 | Jaw movement (2) | speed, acc. & jerk metrics | SIT_7 |
| 7 | Jaw movement (3) | speed, acc. & jerk metrics | SIT_5 |
| 8 | Jaw movement (4) | min + max speed, acc. & jerk metrics | Picture Description |
| 9 | Jaw movement (5) | speed, acc. & jerk metrics | SIT_{9,11,13,15}, Bamboo, Picture Description (mean) |
| 10 | Mouth symmetry | mean mouth symmetry | all |
| 11 | Eye opening | mean and max eye opening | all |

- **Audiovisual data** collected using the Modality multimodal dialog system
- **Tasks** included: (a) **Sentence intelligibility test (SIT)**, (b) **Diadochokinesis (DDK)** involving repetition of syllables, (c) **Read speech task** (Bamboo reading passage), (d) **Picture description task**

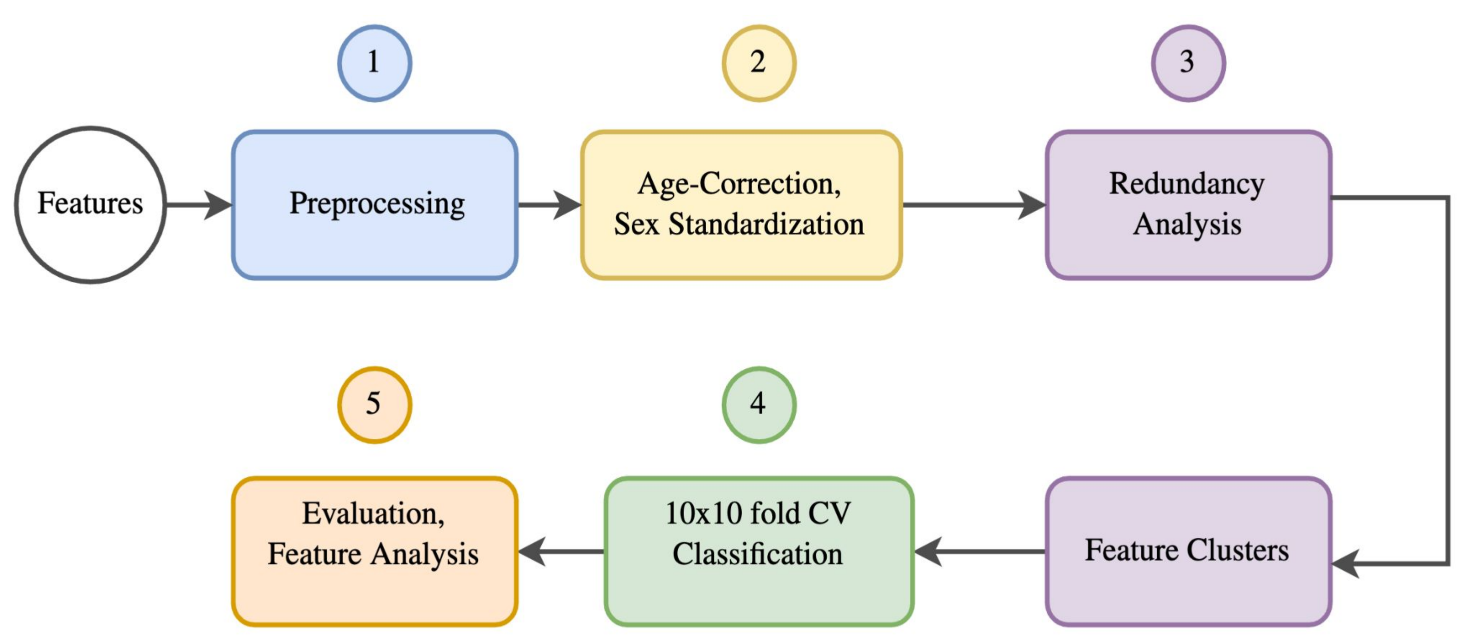


**Figure 2.** Feature Selection & Classification Pipeline.

- See Figure 2 for an overview of the methods applied.
- Additional classification & evaluation details (4+5):
  - **Binary and multi-class classification** using a MLP classifier.
  - **Feature selection** based on highest **effect size per feature cluster** within each fold.
  - Evaluation using F1 score, sensitivity, and specificity.

## Results

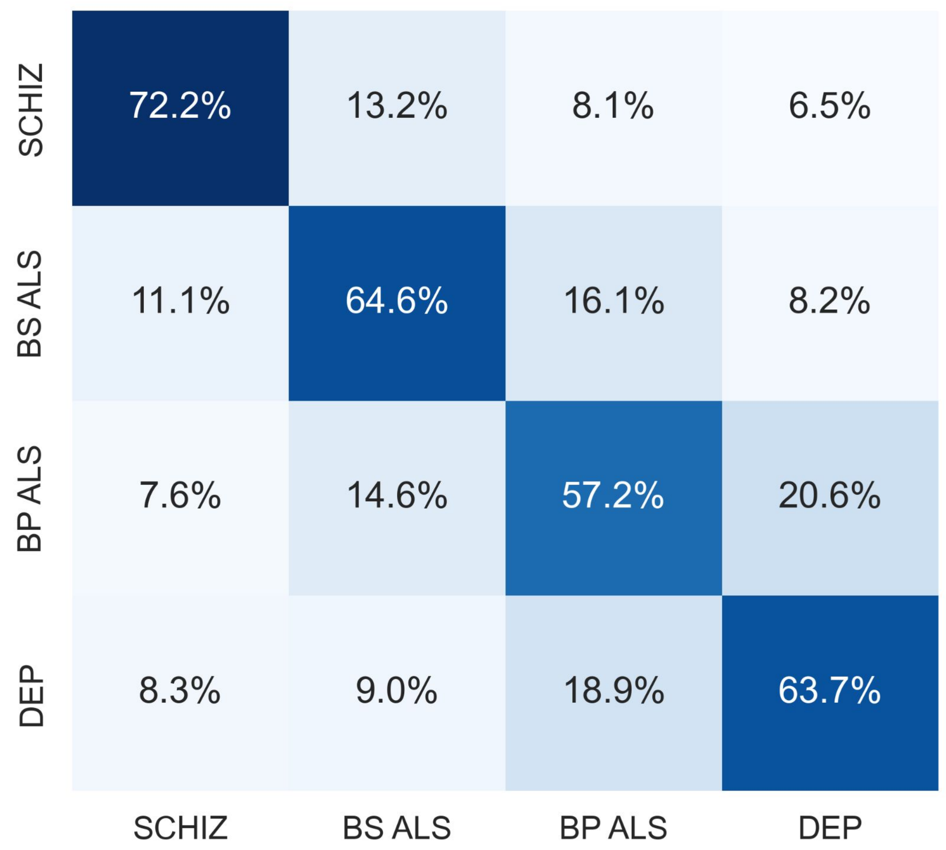


**Figure 3.** 4-class confusion matrix using both speech and facial features. F1-score: 0.64

| Cohort | Speech | Facial | Speech + Facial | | |
|---|---|---|---|---|---|
| | F1 | F1 | F1 | SEN | SP |
| DEP vs. HC | 0.64 | 0.59 | **0.65** | 0.65 | 0.65 |
| SCHIZ vs. HC | 0.82 | 0.64 | **0.83** | 0.85 | 0.82 |
| BP ALS vs. HC | **0.54** | 0.51 | 0.52 | 0.52 | 0.53 |
| BS ALS vs. HC | **0.84** | 0.63 | 0.83 | 0.82 | 0.83 |

**Table 2:** Binary classification (baseline) results.

| Cohort | Speech | Facial | Speech + Facial | | |
|---|---|---|---|---|---|
| | F1 | F1 | F1 | SEN | SP |
| SCHIZ | **0.72** | 0.53 | **0.72** | 0.72 | 0.91 |
| BP ALS | 0.55 | 0.36 | **0.57** | 0.57 | 0.86 |
| BS ALS | 0.62 | 0.47 | **0.64** | 0.65 | 0.88 |
| DEP | 0.61 | 0.46 | **0.64** | 0.64 | 0.88 |
| Average | 0.63 | 0.46 | **0.64** | 0.65 | 0.88 |

**Table 3.** 4-class classification results. SEN: Sensitivity, SP: Specificity

- See Table 2 & 3: Overall, **employing both speech and facial features** is **beneficial** (exception: adding facial information does not enhance performance in binary HC vs. ALS cohorts experiments).
- See Table 1 & 2 + Figure 3:
  - **Best results** in detecting **schizophrenia**, followed by BS ALS in 4-class classification (Binary baseline: approximately equal performance in these disorders).
  - **Greatest challenge** in accurately predicting **BP ALS** (both binary & 4-class).

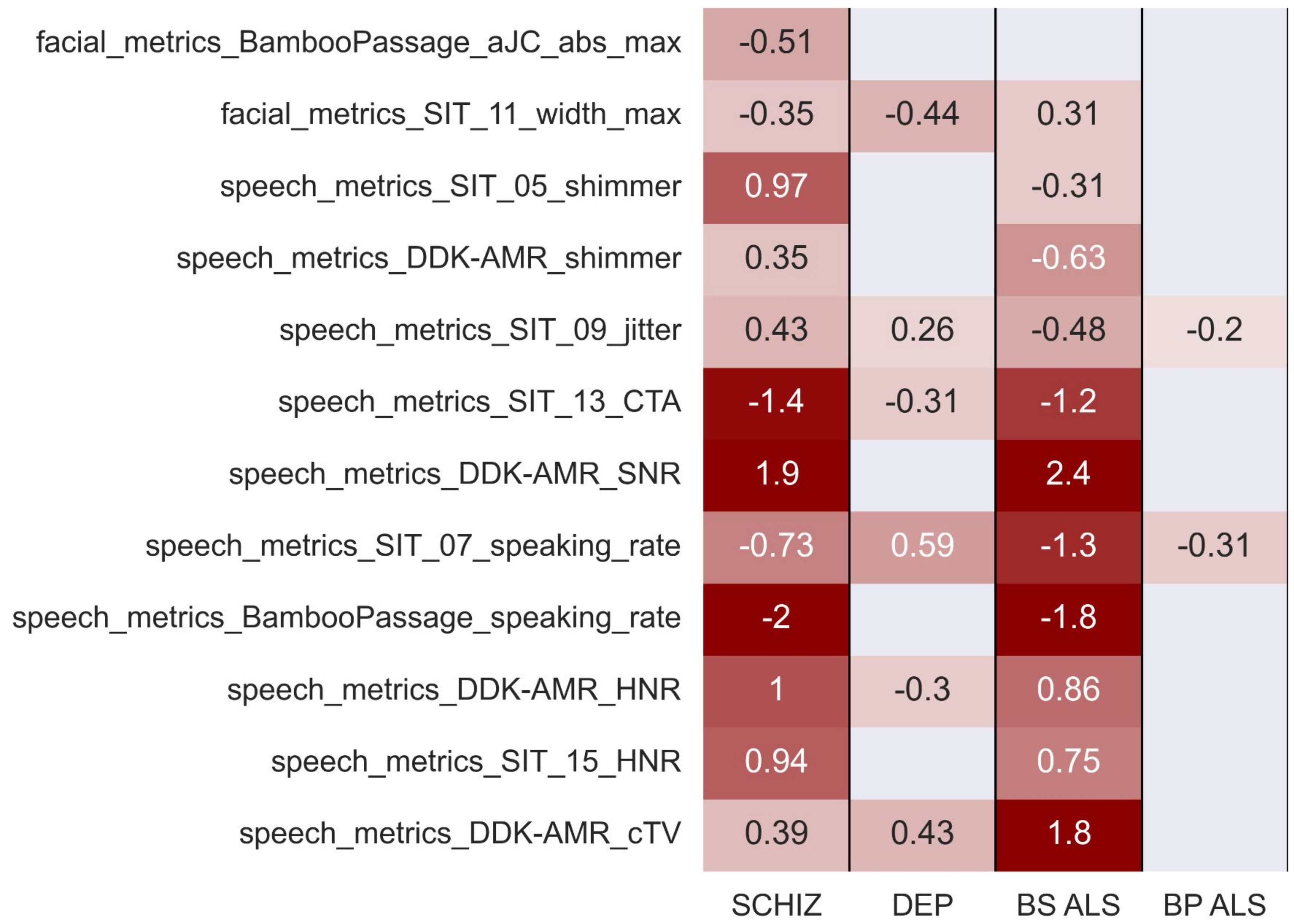


**Figure 4.** Effect Sizes for features selected across all multi-class classification folds. Magnitudes: small: 0.2 - 0.5, medium: 0.5 - 0.8 and large: > 0.8

- See Figure 4: **Consistently chosen features primarily** include **speech** features of timing, voice quality, and energy domains. Additionally, **two facial features stand out**: maximum lip width and maximum absolute acceleration of jaw movements.
- Many features show statistical significance across disorders when compared to controls, while the **effect size magnitudes**, **feature combinations** and in some cases the **direction of effects differ**.

## Conclusions

- **Combining speech and facial information** proved particularly beneficial in the more complex task of multi-classification, with varying accuracy across disorders; **highest separability** shown for **schizophrenia,** lowest for BP ALS.
- **Feature analysis** indicates **several features** that are relevant across experiments (**speech > facial**).
- **Future work: improve generalizability & explainability**, while addressing limitations such as small sample size and lack of information on comorbidities.

This work was funded in part by the National Institutes of Health grant R42DC019877. We thank all study participants for their time and we gratefully acknowledge the contribution of the Peter Cohen Foundation and EverythingALS towards participant recruitment and data collection for the ALS corpus and Anzalee Khan and Jean-Pierre Lindenmayer at the Manhattan Psychiatric Center -- Nathan Kline Institute for the schizophrenia corpus.