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

Vanessa Richter\textsuperscript{1,\dagger}, Michael Neumann\textsuperscript{1} and Vikram Ramanarayanan\textsuperscript{1,2,*}

\textsuperscript{1}Modality.AI, Inc., San Francisco, CA 94105, United States
\textsuperscript{2}University of California, San Francisco, CA 94127, United States

Abstract
Utilizing computer vision and speech signal processing to assess neurological and mental conditions remotely has the potential to help detecting diseases or monitoring their progression earlier and more accurately. Multimodal features have demonstrated usefulness in identifying cases with a disorder from controls across several health conditions. However, challenges arise in distinguishing between specific disorders during the process of differential diagnosis, where shared characteristics among different disorders may complicate accurate classification. Our aim in this study was to evaluate the utility and accuracy of automatically extracted speech and facial features for differentiating between multiple disorders in a multi-class (differential diagnosis) setting using a machine learning classifier. We use datasets comprising people with depression, bulbar and limb onset amyotrophic lateral sclerosis (ALS), and schizophrenia, in addition to healthy controls. The data was collected in a real-world scenario with a multimodal dialog system, where a virtual guide walked participants through a set of tasks that elicit speech and facial behavior. Our study demonstrates the utility of digital speech and facial biomarkers in assessing neurological and mental disorders for differential diagnosis. Furthermore, this research emphasizes the importance of combining information derived from multiple modalities for a more comprehensive understanding and classification of disorders.

Keywords
differential diagnosis, multi-class, mental disorders, neurological disorders, depression, schizophrenia, amyotrophic lateral sclerosis, digital biomarkers, dialog system, speech, facial, multimodal

1. Introduction
One out of eight individuals in the world lives with a mental health disorder, but most people do not have access to effective care.\textsuperscript{1} Moreover, disorders of the nervous system are the second leading cause of death globally [1].

The development of clinically valid digital biomarkers for neurological and mental disorders that can be automatically extracted could significantly improve patients’ lives. This advancement has the potential to assist clinicians in achieving quicker and more reliable diagnoses by providing fast and objective insights into a patient’s state. Note that the idea here is not to replace the clinician, but to provide effective and assistive tools that can help improve his/her efficiency, speed and accuracy.

Many speech and facial features have shown to be useful in differentiating between different mental and neurological disorders and healthy controls (HCs) [2]. However, it remains unclear how distinctly these features characterize a given disorder. For example, percent pause time (PPT) has been found to differ significantly between people with ALS (pALS) and HCs [3] as well as between people with depression symptoms and HCs [4]. Furthermore, a slower speaking rate differentiates pALS [5] as well as people with schizophrenia [6] from HC.

To assess the utility of automatically computed digital biomarkers to capture specific disease attributes despite such shared characteristics, we aim to answer the following questions:

1. How accurately can a machine learning (ML) classifier differentially distinguish between multiple disorders – depression, schizophrenia, bulbar symptomatic ALS and bulbar presymptomatic ALS?
2. Which modalities and features are most useful for this multi-class classification task – overall and with respect to a given disorder – and how does that compare to a binary classification baseline (controls versus cases in each of the investigated health conditions)?

2. Related Work
Recently, digital speech and facial features have been shown to yield statistically significant differences be-
between cases with neurological or mental disorders and healthy controls, exhibit high specificity and sensitivity in discriminatory ability between those groups, or, a high potential for disease progression and treatment effect monitoring [2, 3, 6, 7, 8, 9, 10, 11, 12]. Several studies have evaluated the detection of neurological and mental disorders in multi-class classification settings as compared to binary case-control studies [13, 14, 15]. Altaf et al. [13] introduced an algorithm for Alzheimer’s disease (AD) detection validated on binary classification and multi-class classification of AD, normal and mild cognitive impairment (MCI). Using the bag of visual word approach, the algorithm enhances texture-based features like the gray level co-occurrence matrix. It integrates clinical data, creating a hybrid feature vector from whole magnetic resonance (MR) brain images. They use the Alzheimer’s Disease Neuro-imaging Initiative dataset (ADNI) and achieve 98.4% accuracy in binary AD versus normal classification and 79.8% accuracy in multi-class AD, normal, and MCI classification. Furthermore, Hansen et al. [14] explored the potential of speech patterns as diagnostic markers for multiple neuropsychiatric conditions by examining recordings from 420 participants with major depressive disorder, schizophrenia, autism spectrum disorder, and non-psychiatric controls. Various models were trained and tested for both binary and multi-class classification tasks using speech and text features. While binary classification models exhibited comparable performance to prior research (F1: 0.54–0.92), multi-class classification showed a notable decrease in performance (F1: 0.35–0.75). The study further demonstrates that combining voice- and text-based models enhances overall performance by 9.4% F1 macro, highlighting the potential of a multimodal approach for more accurate neuropsychiatric condition classification. While these studies show the effectiveness of different types of speech- and facial-derived features for assessing psychiatric conditions in differential diagnosis settings, none of them utilized ‘in-the-wild’ data collected remotely from participants devices with a multimodal dialog system.

3. Multimodal Dialog Platform and Data Collection

Audiovisual data was collected using NEMSI (Neurological and Mental health Screening Instrument) [16], a multimodal dialog system for remote health assessments. An overview of the dataset creation process is illustrated in Figure 1. A virtual guide, Tina, led study participants through various tasks that are designed to elicit speech, facial, and motor behaviors. Having an interactive virtual guide to elicit participants’ behavior allows for scalability while providing a natural but controlled and objective interview environment and data collection. Each session starts with a microphone, speaker, and camera check to ensure that the participant has given their device the permission to access camera and microphone, is able to hear the instructions and the captured signal is of adequate quality. After these tests the virtual guide involves participants in a structured conversation that consists of exercises (speaking tasks, open-ended questions, motor abilities) to elicit speech, facial and motor behaviors relevant to the type of disease being studied. In this work, we focus on tasks that were shared across multiple study protocols for different disease conditions: (a) sentence intelligibility test (SIT), (b) diadochokinesis (DDK), (c) read speech, and (d) a picture description task. For (a), participants were asked to read individual SIT sentences of varying lengths (5-15 words2), while (b) required reading a longer passage (Bamboo reading passage, 99 words). To assess DDK skills (c), participants were asked to repeat a pattern of syllables (/pa ta ka/) as fast as they can until they run out of breath and (d) prompted users to describe a scene in a picture that was shown to them on screen. These tasks are inspired by previous work [17, 18, 19].

3.1. Datasets

An overview of the data used in this study is given in Table 1. While some datasets for a disease may be small, there is a subset of tasks that are shared across research studies. Since the data is collected in the same way (remotely with a personal electronic device), we can create a larger dataset for the healthy population across studies to get a more accurate representation of the properties of normative behavior. For the larger dataset of healthy controls, we identify age-related trends as well as collinearity of features. This information is used to correct control as well as patient feature values from

---

2In the remainder of the paper, the different SIT sentence lengths are treated as separate tasks and are denoted as SIT_n, where n is the length in words.
with a diagnosis of schizophrenia at a state psychiatric
facilities in New York, NY. The study was approved by the
Nathan S. Kline Institute for Psychiatric Research and we
obtained written informed consent from all participants
at the time of screening after explaining details of the
study. The assessment of both patients and controls was
overseen by a psychiatrist.

### Table 1

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Gender</th>
<th>Sessions</th>
<th>Mean Age (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls</td>
<td>Female</td>
<td>408 (63%)</td>
<td>655 (62.8%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>240 (37%)</td>
<td>388 (37.2%)</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>648</td>
<td>1043</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>Female</td>
<td>10 (24.4%)</td>
<td>19 (26.4%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>31 (75.6%)</td>
<td>53 (73.6%)</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>41</td>
<td>72</td>
</tr>
<tr>
<td>Depression</td>
<td>Female</td>
<td>66 (79.5%)</td>
<td>76 (79.2%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>17 (20.5%)</td>
<td>20 (20.8%)</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>83</td>
<td>96</td>
</tr>
<tr>
<td>Bulbar Symptomatic ALS</td>
<td>Female</td>
<td>38 (48.1%)</td>
<td>67 (46.2%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>41 (51.9%)</td>
<td>78 (53.8%)</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>79</td>
<td>145</td>
</tr>
<tr>
<td>Bulbar Presymptomatic ALS</td>
<td>Female</td>
<td>31 (50%)</td>
<td>54 (50.5%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>31 (50%)</td>
<td>53 (49.5%)</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>62</td>
<td>107</td>
</tr>
</tbody>
</table>

Table 1: Cohort demographics. SD: standard deviation.

### 3.1.1. Schizophrenia

Schizophrenia is a chronic brain disorder that affects
approximately 24 million or 1 in 300 people (1 in 222
in adults) worldwide. According to the American Psy-
chiatric Association (APA), active schizophrenia may be
characterized by episodes in which the affected individual
cannot distinguish between real and unreal experiences.
Among individuals with schizophrenia, psychiatric and
medical comorbidities such as substance abuse, anxiety
and depression are common [20, 21, 22]. Buckley et al.
pointed out that depression is estimated to affect half of
the patients. These comorbidities, as well as the variation
in symptoms and medications, make the identification of
multimodal biomarkers for schizophrenia a difficult task.
As can be seen in Table 1, we assessed 41 individuals
with a diagnosis of schizophrenia at a state psychiatric
facility in New York, NY. The study was approved by the
Nathan S. Kline Institute for Psychiatric Research and we
obtained written informed consent from all participants
at the time of screening after explaining details of the
study. The assessment of both patients and controls was
overseen by a psychiatrist.

### 3.1.2. Amyotrophic Lateral Sclerosis

ALS is a neurological disease that affects nerve cells in
the brain and spinal cord that control voluntary muscle
movement. The disease is progressive and there is
currently no cure or effective treatment to reverse its
progression. Global estimates of ALS prevalence range
from 1.9 to 6 per 100,000. Studies on ALS found co-
morbid with dementia, parkinsonism and depressive
symptoms [23]. Diekmann et al. [24] found depression
to occur statistically significantly more often in pALS
compared to HC. In addition, Heidari et al. [25] found
in a meta-analysis of 46 eligible studies that the pooled
prevalence of depression among individuals with ALS to
be 34%, with mild, moderate, and severe depression rates
at 29%, 16%, and 8%, respectively.

As shown in Table 1, data from 79 ALS bulbar symp-
tomatic (BS) and 62 ALS bulbar pre-symptomatic (BP)
patients were collected in cooperation with Everythin-
gALS and the Peter Cohen Foundation. In addition to
the assessment of speech and facial behavior, partici-
pants filled out the ALS Functional Rating Scale-revised
(ALSFRS-R), a standard instrument for monitoring the
progression of ALS [26]. The questionnaire comprises
12 questions about physical ability with each function’s
rating ranging from normal function (score 4) to severe
disability (score 0). It includes four scales for different
domains affected by the disorder: bulbar system, fine
and gross motor skills, and respiratory function. The
ALSFRS-R score is the total of the domain sub-scores,
the sum ranging from 0 to 48. For this study, pALS were
stratified into the following sub-cohorts based on their
bulbar subscore: (a) BS ALS with a bulbar subscore < 12
(first three ALSFRS-R questions) and (b) BP ALS with a
bulbar sub-score = 12.

### 3.1.3. Depression

Depression is a common mental health disorder char-
acterized by persistent sadness and lack of interest or
pleasure in previously enjoyable activities. In addition,
fatigue and poor concentration are common. The effects
depression can be long-lasting or recurrent and can
disruptively affect a person’s ability to lead a fulfilling
life. The disorder is one of the most common causes of
disability in the world. One in six people (16.6%) will
experience depression at some point in their lifetime.\

---

1https://www.who.int/news-room/fact-sheets/detail/
schizophrenia, accessed 05/19/2023
2https://www.everythingsals.org/research
3https://www.targetals.org/2022/11/22/epidemiology-of-als-
incidence-prevalence-and-clusters/, accessed 05/19/2023
4https://www.ninds.nih.gov/health-
information/disorders/amyotrophic-lateral-sclerosis-als, accessed
05/19/2023
5https://www.who.int/health-topics/depression, accessed 06/20/2023
6https://www.who.int/patients-families/depression/what-is-
depression, accessed 06/20/2023
A well-established tool for assessing depression is the Patient Health Questionnaire (PHQ)-8 [27]. The PHQ-8 score ranges from 0 to 24 (higher score indicates more severe depression symptoms).

We investigated at least moderately severe depression cases, based on a cutoff of PHQ-8 ≥ 15. The data for this study, including the completion of the PHQ-8 questionnaire, was collected through crowd-sourcing, resulting in a sample of 83 individuals that scored at or above this cutoff. Statistics for this cohort are summarized in Table 1.

4. Methods

Our procedure is divided into the following stages: (1) feature extraction, (2) preprocessing, (3) age-correction and sex-normalization, (4) redundancy and effect size analysis, and finally (5) classification (binary and multi-class) and evaluation.

4.1. Multimodal Metrics Extraction

In this and the following sections, we use the following terminology: Metric denotes a speech or facial metric in general, and Feature denotes a specific combination of a metric extracted from a certain task, e.g. speaking rate for the SIT task.

Both speech and facial metrics were extracted from the audiovisual recordings (overview in Table 2). To extract facial metrics, we used the Mediapipe FaceMesh software10. More specifically, MediaPipe’s Face Detection is based on BlazeFace [28] and determines the (x, y)-coordinates of the face for every frame. Subsequently, 468 facial landmarks are identified using MediaPipe FaceMesh. We selected 14 key landmarks to compute functionals of facial behavior. Distances between landmarks were normalized by dividing them by the intercaruncular distance. In terms of between- as well as within-subject analyses, when the same position relative to the camera cannot be assumed, Roesler et al. [29] found this to be the most reliable method of normalization. More details and a visual depiction of the landmarks used to calculate facial features can be found in [4]. Speech metrics were computed using Praat [30] and cover different domains, such as energy, timing, voice quality and frequency.

4.2. Preprocessing

We applied the following approach to handle missing data, which can occur for a number of reasons, including incomplete sessions, technical issues, or network problems. First, on the session level, we removed participant sessions that had more than 15% missing values. Then, on the feature level, we filtered out features with more than 10% missing values. These thresholds have been determined empirically. After those removal procedures, we impute remaining missing values with mean feature values for the respective cohort in train and test sets separately.

4.3. Age-Correction & Sex-Normalization

Similar to the approach in Falahati et al. [31], we applied a linear correction algorithm to both patient and control data based on age-related changes in the HC cohort. By calculating age trends and coefficients on healthy controls, we aim to obtain the most accurate estimate of purely age-related changes without the confounding effects of disease-related influences. In detail, for each feature, we fit a linear regression model to age as the independent and the feature as the dependent variable, modeling the age-related changes as a linear deviation. This is done separately for males and females to obtain a sex-specific result. Then, the sex-specific regression coefficients are used to correct feature values for age by subtracting the product of coefficient and age from the feature value for each participant. To account for sex-related differences, we applied sex-specific z-scoring to normalize the features. Z-normalization is a methodology that allows for the comparison or compilation of observations of different cohorts [32]. In addition, the normalization process ensures the comparability of features on different scales by centering the feature distributions around zero with a standard deviation of one. First, the dataset to analyze was divided into male and female participants. Then, each feature was normalized within each sex group using z-scoring.

4.4. Redundancy Analysis and Effect Sizes

To identify collinear features and reduce the high-dimensional feature space, we performed hierarchical clustering on the Spearman rank-order correlations using the age-corrected and sex-normalized larger healthy control dataset. We applied the clustering for speech and facial features separately. The clustering procedure is motivated by the approach in Ienco and Meo [33]. It is based on Ward’s method [34], which aims at minimising within-cluster variance. We implemented it using the scikit-learn library11. A dendrogram was plotted to inspect the correlations between features visually and to determine a suitable distance threshold for generating feature clusters. The threshold choice was based on two major factors: (a) balance between speech and facial clusters as we target roughly an equal number to avoid

10https://google.github.io/mediapipe/
<table>
<thead>
<tr>
<th>Domain</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>signal-to-noise ratio (SNR, dB)</td>
</tr>
<tr>
<td></td>
<td>speaking &amp; articulation duration/rate (sec./WPM), percent pause time (PPT, %), canonical timing agreement (CTA, %)</td>
</tr>
<tr>
<td></td>
<td>cycle-to-cycle temporal variability (cTV, sec.), syllable rate (syl./sec.), number of syllables</td>
</tr>
<tr>
<td></td>
<td>shimmer (%), harmonics-to-noise ratio (HNR, dB), jitter (%)</td>
</tr>
<tr>
<td></td>
<td>mean, min, max &amp; standard deviation (stdev) of fundamental frequency (F0, Hz)</td>
</tr>
<tr>
<td>Video</td>
<td>mean, min &amp; max speed/acceleration/jerk of the jaw center (JC)</td>
</tr>
<tr>
<td></td>
<td>mean, min &amp; max speed/acceleration/jerk of the lower lip (LL)</td>
</tr>
<tr>
<td></td>
<td>mean &amp; max lip aperture, lip width, mouth surface area; mean mouth symmetry ratio</td>
</tr>
<tr>
<td></td>
<td>mean &amp; max eye opening</td>
</tr>
</tbody>
</table>

Table 2
Overview of speech and facial metrics.

<table>
<thead>
<tr>
<th>#</th>
<th>Cluster domain</th>
<th>Metrics</th>
<th>Tasks</th>
<th># Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Energy</td>
<td>SNR</td>
<td>all</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>Timing alignment</td>
<td>CTA</td>
<td>all</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Timing, pauses</td>
<td>PPT</td>
<td>all</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Timing, speaking (1)</td>
<td>articulation/speaking duration</td>
<td>Picture Description</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>DDK articulation</td>
<td>SNR, syl.rate, syl.count &amp; cTV</td>
<td>DDK</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Timing, speaking (2)</td>
<td>articulation/speaking rate/time</td>
<td>SIT_{[5,9]}</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>Timing, speaking (3)</td>
<td>articulation/speaking rate/time</td>
<td>SIT_{[7,11,13,15]}, Reading passage</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td>DDK voice quality</td>
<td>HNR, jitter &amp; shimmer</td>
<td>DDK</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>Voice quality (periodicity)</td>
<td>HNR</td>
<td>all except DDK</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>Voice quality (amplitude variation)</td>
<td>shimmer</td>
<td>all except DDK</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>Voice quality (frequency variation)</td>
<td>jitter</td>
<td>all except DDK</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>Frequency (mean, min)</td>
<td>min &amp; mean F0</td>
<td>all</td>
<td>16</td>
</tr>
<tr>
<td>13</td>
<td>Frequency (max, std)</td>
<td>max &amp; std F0</td>
<td>all</td>
<td>16</td>
</tr>
</tbody>
</table>

∑ 113

Table 3
Speech feature clusters identified by hierarchical clustering.

The clusters are used in the feature selection process as described Section 4.5.

Statistical tests to assess the statistical significance, as well as the magnitude and direction of effects for a given comparison, were conducted within classification folds and as part of a post hoc analysis. Effect sizes were calculated using Glass’s Delta [35]. Here, only features showing statistical significance ($p < 0.05$) in the Mann-Whitney U-test (MWU) were considered.

4.5. Classification

For both the binary and multi-class classification experiments, we used a multilayer perceptron (MLP), which was implemented using the scikit-learn library. The MLP has one hidden layer. We experimented with adding more hidden layers, but found that the minimal configuration with only one layer was beneficial in terms of performance. The hidden layer size $h$ was determined dynamically as

$$ h = \frac{f + c}{2} \quad (1) $$

where $f$ is the number of selected features and $c$ the number of classes. The model was trained with a maximum of 10,000 iterations to allow sufficient time for convergence during training. Model training was stopped when the loss or score was not improving by a defined tolerance threshold. Here, we used scikit-learn’s default of $1e-4$. Additionally, the alpha parameter was set to 0.001, controlling the regularization strength to prevent overfitting. The sgd (stochastic gradient descent) solver was used for optimization during training. The batch size was set to auto, enabling the model to determine the appropriate batch size during training. We used the rectified linear unit function as the activation function.
Ten-fold cross-validation was applied for evaluation in order to maximize the utilization of data for both training and testing purposes. To avoid bias towards the majority group, we created datasets that consist of an equal number samples in each disease condition. For each individual participant, we consider, if available, the first two sessions as data points. Because of the equality constraint, the number of data points was limited by the smallest dataset (schizophrenia). This resulted in 72 randomly selected data points per cohort, summing up to a total of 360 data points. The classification experiments are run ten times to smooth out performance variations and obtain more representative results. We split the data using scikit-learn’s StratifiedGroupKFold to make sure that sessions from the same participant are either in the respective training or testing fold. In each fold, we imputed missing values and standardized features by sex using z-scoring. This was done separately for training and test sets.

As a benchmark, we evaluated binary classification performance of models aimed at distinguishing cases with a disorder from controls. Here, for each cluster of collinear features as described in Section 4.4, the one with the highest effect size was selected for the final feature set as input to the classifier. If no feature showed statistically significant differences between cases and controls in a given cluster, no feature was selected. Hence, the number of clusters determines the maximum number of features fed into the classifier. Statistical significance and effect sizes for each feature were calculated as described in the previous section.

In a second step, we performed 4-class classification, incorporating all the investigated neurological and mental disorders. Here, feature selection was done based on pairwise comparisons of all disease cohorts (e.g. Depression vs. Schizophrenia cases, Schizophrenia vs. BS ALS cases, BS ALS vs. Depression cases, and so on). We merged the selected features from those comparisons as input to the classifier. Therefore, multiple features from the same cluster could be included in one feature set. We allowed a certain amount of redundancy compared to the case-control baseline in order to account for the complexity associated with multiple comparisons. For both experiments, classification performance was evaluated in terms of F1 score, sensitivity, and specificity.

### 5. Results

#### 5.1. Binary Classification Baseline

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Speech F1</th>
<th>Facial F1</th>
<th>Speech + Facial F1</th>
<th>SEN</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEP vs. HC</td>
<td>0.64</td>
<td>0.59</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>SCHIZ vs. HC</td>
<td>0.62</td>
<td>0.64</td>
<td>0.83</td>
<td>0.85</td>
<td>0.82</td>
</tr>
<tr>
<td>BP ALS vs. HC</td>
<td>0.54</td>
<td>0.51</td>
<td>0.52</td>
<td>0.52</td>
<td>0.53</td>
</tr>
<tr>
<td>BS ALS vs. HC</td>
<td>0.84</td>
<td>0.63</td>
<td>0.83</td>
<td>0.82</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table 5

Binary classification results. In each row, we highlighted the highest performance in terms of F1.

HC: Healthy Controls, DEP: Depression, SCHIZ: Schizophrenia, SEN: Sensitivity, SP: Specificity

As can be seen in Table 5, we observe a good performance in classifying controls versus BS ALS (speech features alone; F1-score: 0.84) and schizophrenia (combined speech and facial; F1-score: 0.83) cases, respectively. The binary classification of depression did not perform as well; however, it still surpassed the random chance baseline (combined speech and facial; F1-score: 0.65). The classifier struggled to distinguish controls from BP ALS cases, where we observed performance just above
random chance across modalities. Furthermore, the performance with regard to sensitivity and specificity is relatively balanced across comparisons.

In depression and schizophrenia, combining speech and facial modalities resulted in improved classification performance compared to speech or facial features alone, as shown in Table 5. However, adding facial information did not enhance performance for BP or BS and ALS cohorts compared to utilizing speech features alone.

### 5.2. Multi-Class Classification

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Speech F1</th>
<th>Facial F1</th>
<th>Speech + Facial SEN</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCHIZ</td>
<td>0.72</td>
<td>0.53</td>
<td>0.72 0.72</td>
<td>0.91</td>
</tr>
<tr>
<td>BP ALS</td>
<td>0.55</td>
<td>0.36</td>
<td>0.57 0.57</td>
<td>0.86</td>
</tr>
<tr>
<td>BS ALS</td>
<td>0.62</td>
<td>0.47</td>
<td>0.64 0.65</td>
<td>0.88</td>
</tr>
<tr>
<td>DEP</td>
<td>0.61</td>
<td>0.46</td>
<td>0.64 0.64</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 6
Multi-class classification results. In each row, we highlight the highest F1 score performance.

HC: Healthy Controls, DEP: Depression, SCHIZ: Schizophrenia, SEN: Sensitivity, SP: Specificity

Figure 2: Normalized confusion matrix for 4-class classification. The x-axis shows the true labels, the y-axis the predicted ones.

In the 4-class experiment aimed at discriminating between all investigated neurological and mental disorders, we achieve the best overall performance (F1-score: 0.64) by utilizing both speech and facial features, as shown in Table 6. Overall, the specificity (average: 0.88) for the disorders examined is considerably higher than the sensitivity (average: 0.65). This indicates that the classifier is more effective at avoiding false-positive results than identifying true positives. In most cases, namely for BS ALS, BP ALS and depression the per class F1-score is highest when combining speech and facial features. There is no performance difference between using only speech or speech and facial features for identifying schizophrenia. Figure 2 shows a confusion matrix that indicates the percentage of accurate class predictions and the classes with which they were confused. The model was most confident in detecting schizophrenia (72.22%), followed by BS ALS (64.58%) and depression (63.75%). The model faced its greatest challenge in accurately predicting BP ALS (57.22%), yet it still performs notably above chance in a 4-class classification scenario. BP ALS and depression cases were most often confused with each other. Schizophrenic patients were least often confused with other cohorts. Among the cases of BS ALS, the most frequent confusion occurred with BP ALS patients (16.11%).

The features that we identified to be consistently chosen across classification folds (Table 7) are predominantly speech features of timing, voice quality, and energy domains. In addition, two facial features are selected across folds concerning the maximum lip width and the maximum absolute acceleration of jaw movements. We conducted a post hoc analysis of effect sizes between HC and cases with a disorder for these features to gain further insight into disorder-specific importance. Here, positive effect sizes represent feature values that are larger for cases with a disorder than controls. Conversely, negative values represent larger feature values for controls than cases with a disorder. In schizophrenia, we find all of the features consistently selected across classification folds to be statistically significant when compared to HC. With respect to the other cohorts, the largest effects are shown for CTA (-1.44 for SIT_13) and speaking rate (-2.00 for RP). This shows that patients exhibit a lower CTA, a measure of phonetic alignment between their own speech and that of the virtual guide, while speaking slower. We also observed a smaller average lip width as an important feature that shows the largest effect between HC and depression cases compared to the other cohorts. This may be associated with decreased emotional expressivity, as indicated by reduced smiling and increased frowning. These findings align with previous studies highlighting similar patterns of emotional expressiveness in depression [37, 38]. Few and small differences compared to controls are revealed for BP ALS cases. This is also the cohort with the lowest performance across classification experiments. In BS ALS, we found the largest effects for SNR and speaking rate. Another feature that stood out is cTV in the DDK task, a measure that captures the temporal variability, i.e. the consistency or irregularity in the timing of speech patterns, between consecutive cycles of speech.

---

12We follow the commonly used effect size magnitude thresholds as suggested in Cohen [36] – small: 0.2 – 0.5, medium: 0.5 – 0.8, and large: > 0.8
Table 7

Features selected across all multi-class classification CV folds (considering the 4 disorders) and their effect sizes as calculated between the healthy control and disorder cohorts. In each row, we highlighted the largest effect size, which were only calculated in case of statistical significance.

<table>
<thead>
<tr>
<th>Features</th>
<th>Modality</th>
<th>Cluster domain</th>
<th>SCHIZ</th>
<th>BP ALS</th>
<th>BS ALS</th>
<th>DEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>max abs acc. JC (RP)</td>
<td>Facial</td>
<td>Jaw movement</td>
<td>-0.51</td>
<td></td>
<td>N.S</td>
<td></td>
</tr>
<tr>
<td>max lip width (SIT 11)</td>
<td>Facial</td>
<td>Lip width</td>
<td>-0.35</td>
<td>0.31</td>
<td>-0.44</td>
<td></td>
</tr>
<tr>
<td>shimmer (DDK)</td>
<td>Speech</td>
<td>Voice quality</td>
<td>0.35</td>
<td>-0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shimmer (SIT 5)</td>
<td>Speech</td>
<td>Voice quality</td>
<td>0.97</td>
<td>-0.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>jitter (SIT 9)</td>
<td>Speech</td>
<td>Voice quality</td>
<td>0.43</td>
<td>-0.20</td>
<td>-0.48</td>
<td>0.26</td>
</tr>
<tr>
<td>CTA (SIT 13)</td>
<td>Speech</td>
<td>Timing alignment</td>
<td>-1.44</td>
<td>-1.16</td>
<td>-0.31</td>
<td></td>
</tr>
<tr>
<td>SNR (DDK)</td>
<td>Speech</td>
<td>Energy</td>
<td>1.88</td>
<td>-2.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>speaking rate (RP)</td>
<td>Speech</td>
<td>Timing, speaking</td>
<td>-2.00</td>
<td>-1.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>speaking rate (SIT 7)</td>
<td>Speech</td>
<td>Timing, speaking</td>
<td>-0.73</td>
<td>-0.31</td>
<td>1.25</td>
<td>0.59</td>
</tr>
<tr>
<td>HNR (DDK)</td>
<td>Speech</td>
<td>Voice quality</td>
<td>1.01</td>
<td>0.86</td>
<td>-0.30</td>
<td></td>
</tr>
<tr>
<td>HNR (SIT 15)</td>
<td>Speech</td>
<td>Voice quality</td>
<td>0.94</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cTV (DDK)</td>
<td>Speech</td>
<td>Energy &amp; articulation skills</td>
<td>0.39</td>
<td>1.82</td>
<td>0.43</td>
<td></td>
</tr>
</tbody>
</table>

While many features are shared in terms of indicating a signal between cases with a disorder and controls, it is mostly the magnitude of the effect that differentiates them, as well as how they combine. However, there are also a few features that show a different direction of effect across cohorts. For example, in BS ALS, compared to other cohorts, we observed the largest effect for shimmer (DDK, -0.63), which measures the variation in amplitude of the vocal folds during the speech signal. There is no effect observed for BP ALS or depression cohorts, while in schizophrenia, the direction of effect is the opposite (0.35).

6. Discussion

We explored the utility of speech and facial features extracted by a multimodal dialog system for differential classification of ALS, depression and schizophrenia. Note that the idea here is not to replace clinicians, but to provide effective and assistive tools that can help improve their efficiency, speed and accuracy. Overall, combining speech and facial information proved to be beneficial for identifying several disorders in both multi-class and binary classification experiments. In addition, our automated feature analysis indicates several features that show relevance across experiments. While some of these features are intuitively identifiable by human experts as markers of a given disorder (for example, a slower speaking rate or a lower intelligibility), such an analysis also allows discovery of other features that might be harder to detect or identify objectively by human experts, such as quicker facial movements.

That being said, we acknowledge the importance of contextualizing the promise of such multimodal methodologies for differential diagnosis with several caveats. First, the performance of any machine learning classifier trained for this purpose will depend on the specific conditions being studied and the range and heterogeneity of symptoms presented in each case. For example, in this study we investigated four specific conditions – schizophrenia, depression, bulbar symptomatic (BS) and bulbar presymptomatic (BP) ALS – and we observed that schizophrenia (where the facial modality is particularly good at capturing characteristics exhibited therein such as anhedonia, blunted affect, etc.) and BS ALS (which is characterized by speech motor deficits, reflected in the timing, rate and intelligibility of speech), quite different in terms of symptom presentation, exhibit greater separability relative to other classes for differential classification. For both BS ALS and schizophrenia, our analysis demonstrates a robust discriminatory capability to effectively distinguish these cohorts from healthy controls, as well as other neurological and mental disorders, in binary and multi-class experiments. However, the overall higher specificity of the multi-class classifier implies a robust capability to accurately identify non-cases, effectively minimizing false positives. Yet, the lower sensitivity suggests limitations in the identification of true cases for the analyzed disorders, likely due to the imposed strong restrictions. In BS ALS, speech features alone demonstrate superior performance when comparing this group with controls. Yet, in the more intricate task of differential diagnosis, performance improves when speech features are combined with facial information. For schizophrenia, the combination of speech and facial modalities proves most effective in both binary and multitask experiments.
In contrast, BP ALS, which does not present with as many speech and facial motor deficits, is much less separable even in binary classification, let alone in the multi-class classification context, highlighting the challenging nature of detecting this condition. Furthermore, for the misidentified BS ALS cases, the classifier most frequently categorized them as BP ALS. Although distinguishing BP ALS cases from controls is challenging, this outcome indicates that the classifier may have only been able to capture condition-specific information from features that are shared across different stages of ALS, which may have led to this confusion. Finally, in evaluating depression, best performance in both binary and multi-class classification experiments is achieved by combining speech and facial information. The overall accuracy in discerning depression from other cohorts is notably lower compared to schizophrenia or BS ALS. The variability introduced by the wide range and time horizon of potential symptoms present in depression as well as medication status might contribute to lower differential diagnosis accuracy. That being said, a significant limitation of the present study is the lack of information about co-morbidities to factor into our analysis, since datasets were collected independently. Future research will aim to explicitly address this gap by capturing, for instance, information about co-morbid depression in ALS or schizophrenia (e.g., through PHQ-8 scales), that might help us better stratify these cohorts.

Second, this study focused on a restricted set of tasks, primarily focusing on reading abilities and picture description assessments. However, these task-feature combinations alone may not fully capture the nuances of each disorder.

Third, while we focused on interpretable features in this study, less interpretable ones, such as log mel spectrograms or Mel Frequency Cepstral Coefficients (MFCCs) may be able to capture more nuanced and complex patterns in the data. Additionally, more sophisticated deep learning approaches for representation learning could be applied, such as Res-Net 50 [39] in the facial modalities. While such features can be powerful in capturing subtle details and nuances of audiovisual behavior, the inner workings of the deep learning model are not easily explainable or interpretable by non-experts.

Fourth, our sample size is not representative enough to truly claim generalizability of findings. The smaller the sample, the larger the risk of having model "blind spots" that in turn lead to variable estimates of true model performance. That being said, our results argue for the importance of a hybrid approach to differential diagnosis going forward, combining knowledge-driven and data-driven approaches. Understanding specific disease pathologies and symptoms can in turn help in developing features and learning methodologies that lead to better separability of disease conditions. Future work will also focus on improving differential diagnosis performance in a manner that is both generalizable and explainable.

Acknowledgments
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.

References


