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 94105, United States
[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.[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-

Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI 2024, Vancouver, BC, Canada
*Corresponding author.
†Vanessa Richter performed the work described in this paper when she was an intern at Modality.AI.
vikram.ramanarayanan@modality.ai (V. Ramanarayanan)

[1]https://www.who.int/news-room/fact-sheets/detail/mental-disorders, accessed 11/7/2022

tween 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

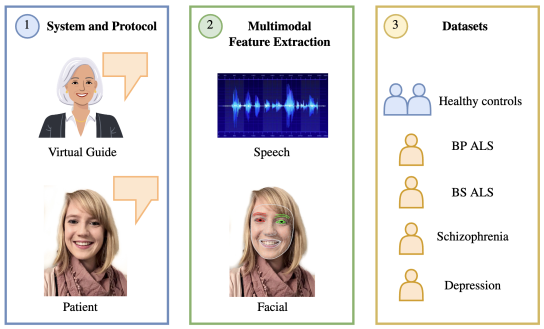

Figure 1: Overview of feature extraction and dataset creation.

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 words[2]), 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

[2]In 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.

	Participants	Sessions	Mean Age (SD)
Controls			
Female	408 (63%)	655 (62.8%)	46.3 (16.4)
Male	240 (37%)	388 (37.2%)	46.2 (16.0)
All	648	1043	46.3 (16.2)
Schizophrenia			
Female	10 (24.4%)	19 (26.4%)	36.1 (9.4)
Male	31 (75.6%)	53 (73.6%)	36.6 (10.1)
All	41	72	36.5 (9.9)
Depression			
Female	66 (79.5%)	76 (79.2%)	34.6 (12.1)
Male	17 (20.5%)	20 (20.8%)	35.0 (10.2)
All	83	96	34.7 (11.7)
Bulbar Symptomatic ALS			
Female	38 (48.1%)	67 (46.2%)	61.7 (10.8)
Male	41 (51.9%)	78 (53.8%)	61.3 (9.0)
All	79	145	61.5 (9.8)
Bulbar Presymptomatic ALS			
Female	31 (50%)	54 (50.5%)	58.1 (10.9)
Male	31 (50%)	53 (49.5%)	62.2 (8.3)
All	62	107	60.1 (9.9)

Table 1
Cohort demographics. SD: standard deviation.

age effects and remove feature redundancies.

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)[3] worldwide. According to the American Psychiatric Association (APA), active schizophrenia may be characterized by episodes in which the affected individual cannot distinguish between real and unreal experiences.[4] 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] https://www.who.int/news-room/fact-sheets/detail/schizophrenia, accessed 05/19/2023
[4] https://www.psychiatry.org/patients-families/schizophrenia/what-is-schizophrenia, accessed 05/19/2023

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.[5]. Global estimates of ALS prevalence range from 1.9 to 6 per 100,000.[6] Studies on ALS found comorbidity 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 symptomatic (BS) and 62 ALS bulbar pre-symptomatic (BP) patients were collected in cooperation with EverythingALS and the Peter Cohen Foundation[7]. In addition to the assessment of speech and facial behavior, participants 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 characterized by persistent sadness and lack of interest or pleasure in previously enjoyable activities. In addition, fatigue and poor concentration are common. The effects of depression can be long-lasting or recurrent and can drastically affect a person's ability to lead a fulfilling life. The disorder is one of the most common causes of disability in the world.[8] One in six people (16.6%) will experience depression at some point in their lifetime.[9]

[5] https://www.ninds.nih.gov/health-information/disorders/amyotrophic-lateral-sclerosis-als, accessed 05/19/2023
[6] https://www.targetals.org/2022/11/22/epidemiology-of-als-incidence-prevalence-and-clusters/, accessed 05/19/2023
[7] https://www.everythingals.org/research
[8] https://www.who.int/health-topics/depression, accessed 06/20/2023
[9] https://www.psychiatry.org/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 $\geq$ 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 software[10]. 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 features. 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` library[11]. 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

[10] https://google.github.io/mediapipe/

[11] https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html

	Domain	Metrics
Audio	Energy	signal-to-noise ratio (SNR, dB)
	Timing	speaking & articulation duration/rate (sec./WPM), percent pause time (PPT, %), canonical timing agreement (CTA, %)
	Specific to DDK	cycle-to-cycle temporal variability (cTV, sec.), syllable rate (syl./sec.), number of syllables
	Voice quality	shimmer (%), harmonics-to-noise ratio (HNR, dB), jitter (%)
	Frequency	mean, min, max & standard deviation (stdev) of fundamental frequency (F0, Hz)
Video	Jaw	mean, min & max speed/acceleration/jerk of the jaw center (JC)
	Lower Lip	mean, min & max speed/acceleration/jerk of the lower lip (LL)
	Mouth	mean & max lip aperture, lip width, mouth surface area; mean mouth symmetry ratio
	Eyes	mean & max eye opening

Table 2
Overview of speech and facial metrics.

#	Cluster domain	Metrics	Tasks	# Features
1	Energy	SNR	all	8
2	Timing alignment	CTA	all	6
3	Timing, pauses	PPT	all	5
4	Timing, speaking (1)	articulation/speaking duration	Picture Description	2
5	DDK articulation	SNR, syl.rate, syl.count & cTV	DDK	4
6	Timing, speaking (2)	articulation/speaking rate/time	SIT_{5,9}	8
7	Timing, speaking (3)	articulation/speaking rate/time	SIT_{7,11,13,15}, Reading passage	21
8	DDK voice quality	HNR, jitter & shimmer	DDK	3
9	Voice quality (periodicity)	HNR	all except DDK	8
10	Voice quality (amplitude variation)	shimmer	all except DDK	8
11	Voice quality (frequency variation)	jitter	all except DDK	8
12	Frequency (mean, min)	min & mean F0	all	16
13	Frequency (max, std)	max & std F0	all	16
			$\sum$	113

Table 3
Speech feature clusters identified by hierarchical clustering.

predominance of one modality over the other, and (b) expert knowledge about the different task and feature domains (e.g. timing versus voice quality features, jaw versus eye movement or read versus free speech), which resulted in the clusters shown in Table 3 and Table 4. 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

$$h = \frac{f + c}{2} \tag{1}$$

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.

#	Cluster domain	Metrics	Tasks	# Features
1	Lip movement (1)	speed, acc. & jerk measures	all except DDK	95
2	Lip width	mean & max lip width	all	18
3	Mouth opening	mean & max lip aperture, mouth surface area	all	36
4	Lip movement (2)	speed, acc. & jerk metrics	DDK	12
5	Jaw movement (1)	speed, acc. & jerk metrics	DDK	12
6	Jaw movement (2)	speed, acc. & jerk metrics	SIT_7	12
7	Jaw movement (3)	speed, acc. & jerk metrics	SIT_5	12
8	Jaw movement (4)	min + max speed, acc. & jerk metrics	Picture Description	9
9	Jaw movement (5)	speed, acc. & jerk metrics	SIT_{9,11,13,15}, RP, Picture Description	63
10	Mouth symmetry	mean mouth symmetry	all	9
11	Eye opening	mean and max eye opening	all	18
			$\sum$	296

Table 4
Facial feature clusters identified by hierarchical clustering. RP: reading passage.

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 these 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

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 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

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 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

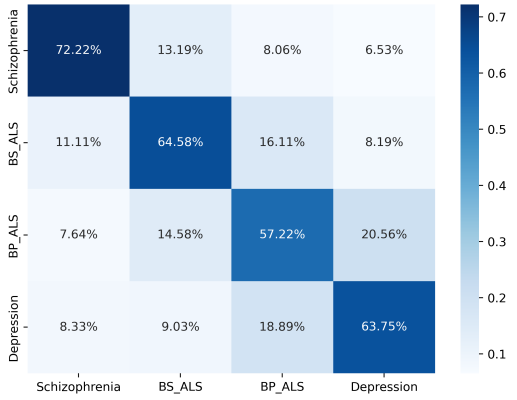

Figure 2: Normalized confusion matrix for 4-class classification. The x-axis shows the true labels, the y-axis the predicted ones.

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[12]. 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.

[12]We 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

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

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.
HC: healthy controls, SCHIZ: schizophrenia, BS: bulbar symptomatic, BP: bulbar pre-symptomatic, DEP: depression, JC: jaw center, RP: reading passage

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 be 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 modality. 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 on unseen real world data, giving algorithm designers an inaccurate sense of how well a model is performing during development [40].

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

[1] V. Feigin, E. Nichols, T. Alam, M. Bannick, E. Beghi, N. Blake, W. Culpepper, E. Dorsey, A. Elbaz, R. Ellenbogen, J. Fisher, C. Fitzmaurice, G. Giussani, L. Glennie, S. James, C. Johnson, N. Kassebaum, G. Logroscino, B. Marin, T. Vos, Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the global burden of disease study 2016, The Lancet Neurology 18 (2019) 459–480. doi:10.1016/S1474-4422(18)30499-X.

[2] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, J. R. Green, Speech as a biomarker: Opportunities, interpretability, and challenges, Perspectives of the ASHA Special Interest Groups 7 (2022) 276–283.

[3] M. Neumann, O. Roesler, J. Liscombe, H. Kothare, D. Suendermann-Oeft, J. D. Berry, E. Fraenkel, R. Norel, A. Anvar, I. Navar, A. V. Sherman, J. R. Green, V. Ramanarayanan, Multimodal dialog based speech and facial biomarkers capture differential disease progression rates for als remote patient monitoring, in: Proceedings of the 32nd International Symposium on Amyotrophic Lateral Sclerosis and Motor Neuron Disease, Virtual, 2021.

[4] V. Richter, J. Cohen, M. Neumann, D. Black, A. Haq, J. Wright-Berryman, V. Ramanarayanan, A multimodal dialog approach to mental state characterization in clinically depressed, anxious, and suicidal populations, Frontiers in Psychology 14 (2023). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1135469. doi:10.3389/fpsyg.2023.1135469.

[5] M. Neumann, O. Roesler, J. Liscombe, H. Kothare, D. Suendermann-Oeft, D. Pautler, I. Navar, A. Anvar, J. Kumm, R. Norel, E. Fraenkel, A. Sherman, J. Berry, G. Pattee, J. Wang, J. Green, V. Ramanarayanan,

Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale, 2021, pp. 4783–4787. doi:10.21437/Interspeech.2021-1801.

[6] V. Richter, M. Neumann, H. Kothare, O. Roesler, J. Liscombe, D. Suendermann-Oeft, S. Prokop, A. Khan, C. Yavorsky, J.-P. Lindenmayer, V. Ramanarayanan, Towards multimodal dialog-based speech & facial biomarkers of schizophrenia, in: Companion Publication of the 2022 International Conference on Multimodal Interaction, ICMI '22 Companion, Association for Computing Machinery, New York, NY, USA, 2022, p. 171–176. URL: https://doi.org/10.1145/3536220.3558075. doi:10.1145/3536220.3558075.

[7] H. Kothare, M. Neumann, J. Liscombe, O. Roesler, W. Burke, A. Exner, S. Snyder, A. Cornish, D. Habberstad, D. Pautler, D. Suendermann-Oeft, J. Huber, V. Ramanarayanan, Statistical and clinical utility of multimodal dialogue-based speech and facial metrics for parkinson's disease assessment, 2022, pp. 3658–3662. doi:10.21437/Interspeech.2022-11048.

[8] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, J. Epps, Diagnosis of depression by behavioural signals: A multimodal approach, in: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 11–20. URL: https://doi.org/10.1145/2512530.2512535. doi:10.1145/2512530.2512535.

[9] J. Robin, M. Xu, A. Balagopalan, J. Novikova, L. Kahn, A. Oday, M. Hejrati, S. Hashemifar, M. Negahdar, W. Simpson, E. Teng, Automated detection of progressive speech changes in early alzheimer's disease, Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 15 (2023) e12445. doi:https://doi.org/10.1002/dad2.12445.

[10] J. Hlavnika, R. Cmejla, T. Tykalová, K. onka, E. Růika, J. Rusz, Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder, Scientific Reports 7 (2017). URL: https://api.semanticscholar.org/CorpusID:19272861.

[11] G. Stegmann, S. Charles, J. Liss, J. Shefner, S. Rutkove, V. Berisha, A speech-based prognostic model for dysarthria progression in als, Amyotrophic lateral sclerosis & frontotemporal degeneration (2023) 1–6. URL: https://doi.org/10.1080/21678421.2023.2222144. doi:10.1080/21678421.2023.2222144, advance online publication.

[12] J. R. Green, K. M. Allison, C. Cordella, B. D. Richburg, G. L. Pattee, J. D. Berry, E. A. Macklin, E. P. Pioro, R. A. Smith, Additional evidence for a therapeutic effect of dextromethorphan/quinidine on bulbar motor function in patients with amyotrophic lateral sclerosis: A quantitative speech analysis, British Journal of Clinical Pharmacology 84 (2018) 2849–2856.

[13] T. Altaf, S. M. Anwar, N. Gul, M. N. Majeed, M. Majid, Multi-class alzheimer's disease classification using image and clinical features, Biomedical Signal Processing and Control 43 (2018) 64–74. URL: https://www.sciencedirect.com/science/article/pii/S1746809418300508. doi:https://doi.org/10.1016/j.bspc.2018.02.019.

[14] L. Hansen, R. Rocca, A. Simonsen, et al., Speech- and text-based classification of neuropsychiatric conditions in a multidiagnostic setting, Nature Mental Health (2023). doi:10.1038/s44220-023-00152-7.

[15] E. Emre, Erol, C. Taş, N. Tarhan, Multi-class classification model for psychiatric disorder discrimination, International Journal of Medical Informatics 170 (2023) 104926. URL: https://www.sciencedirect.com/science/article/pii/S1386505622002404. doi:https://doi.org/10.1016/j.ijmedinf.2022.104926.

[16] D. Suendermann-Oeft, A. Robinson, A. Cornish, D. Habberstad, D. Pautler, D. Schnelle-Walka, F. Haller, J. Liscombe, M. Neumann, M. Merrill, O. Roesler, R. Geffarth, Nemsi: A multimodal dialog system for screening of neurological or mental conditions, in: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, IVA '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 245–247. URL: https://doi.org/10.1145/3308532.3329415. doi:10.1145/3308532.3329415.

[17] A. K. Silbergleit, A. F. Johnson, B. H. Jacobson, Acoustic analysis of voice in individuals with amyotrophic lateral sclerosis and perceptually normal vocal quality, Journal of Voice 11 (1997) 222–231.

[18] B. Tomik, R. J. Guiloff, Dysarthria in amyotrophic lateral sclerosis: A review, Amyotrophic Lateral Sclerosis 11 (2010) 4–15.

[19] M. Novotny, J. Melechovsky, K. Rozenstoks, T. Tykalova, P. Kryze, M. Kanok, J. Klempir, J. Rusz, Comparison of automated acoustic methods for oral diadochokinesis assessment in amyotrophic lateral sclerosis, Journal of speech, language, and hearing research : JSLHR 63 (2020) 3453–3460. doi:10.1044/2020_JSLHR-20-00109.

[20] P. Buckley, B. Miller, D. Lehrer, D. Castle, Psychiatric comorbidities and schizophrenia, Schizophrenia bulletin 35 (2008) 383–402. doi:10.1093/schbul/sbn135.

[21] A. I. Green, C. M. Canuso, M. J. Brenner, J. D. Wojcik, Detection and management of comorbidity in patients with schizophrenia, Psychiatric Clinics 26 (2003) 115–139.

[22] G. B. Cassano, S. Pini, M. Saettoni, P. Rucci, L. Dell'Osso, Occurrence and clinical correlates of psychiatric comorbidity in patients with psychotic disorders, Journal of Clinical Psychiatry 59 (1998) 60–68.

[23] S. Körner, K. Kollewe, J. Ilsemann, A. Karch, R. Dengler, K. Krampfl, S. Petri, Prevalence and prognostic impact of comorbidities in amyotrophic lateral sclerosis, European journal of neurology : the official journal of the European Federation of Neurological Societies 20 (2012). doi:10.1111/ene.12015.

[24] K. Diekmann, M. Kuźma-Kozakiewicz, M. Piotrkiewicz, M. Gromicho, J. Grosskreutz, P. M. Andersen, M. de carvalho, H. Uysal, A. Osmanovic, O. Schreiber-Katz, S. Petri, S. Körner, Impact of comorbidities and co-medication on disease onset and progression in a large german als patient group, Journal of Neurology 267 (2020). doi:10.1007/s00415-020-09799-z.

[25] M. E. Heidari, J. Nadali, A. Parouhan, M. Azarafraz, S. M. Tabatabai, S. S. N. Irvani, F. Eskandari, A. Gharebaghi, Prevalence of depression among amyotrophic lateral sclerosis (als) patients: A systematic review and meta-analysis, Journal of affective disorders 287 (2021) 182–190. doi:10.1016/j.jad.2021.03.015.

[26] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function, Journal of the Neurological Sciences 169 (1999) 13–21. URL: https://www.sciencedirect.com/science/article/pii/S0022510X99002105. doi:https://doi.org/10.1016/S0022-510X(99)00210-5.

[27] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, A. H. Mokdad, The phq-8 as a measure of current depression in the general population, Journal of Affective Disorders 114 (2009) 163–173. URL: https://www.sciencedirect.com/science/article/pii/S0165032708002826. doi:https://doi.org/10.1016/j.jad.2008.06.026.

[28] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, M. Grundmann, Blazeface: Sub-millisecond neural face detection on mobile gpus, CoRR abs/1907.05047 (2019). URL: http://arxiv.org/abs/1907.05047. arXiv:1907.05047.

[29] O. Roesler, H. Kothare, W. Burke, M. Neumann, J. Liscombe, A. Cornish, D. Habberstad, D. Pautler, D. Suendermann-Oeft, V. Ramanarayanan, Exploring facial metric normalization for within- and between-subject comparisons in a multimodal health monitoring agent, in: Companion Publication of the 2022 International Conference on Multimodal Interaction, ICMI '22 Companion, Association for Computing Machinery, New York, NY, USA, 2022, p. 160–165. URL: https://doi.org/10.1145/3536220.3558071. doi:10.1145/3536220.3558071.

[30] P. Boersma, V. Van Heuven, Speak and unspeak with praat, Glot International 5 (2001) 341–347.

[31] F. Falahati, D. Ferreira, J.-S. Muehlboeck, H. Soininen, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, C. Spenger, S. Lovestone, M. Eriksdotter, L.-O. Wahlund, A. Simmons, E. Westman, The effect of age correction on multivariate classification in alzheimer's disease, with a focus on the characteristics of incorrectly and correctly classified subjects, Brain Topography In-press (2016). doi:10.1007/s10548-015-0455-1.

[32] J.-P. Guilloux, M. Seney, N. Edgar, E. Sibille, Integrated behavioral z-scoring increases the sensitivity and reliability of behavioral phenotyping in mice: Relevance to emotionality and sex, Journal of neuroscience methods 197 (2011) 21–31. doi:10.1016/j.jneumeth.2011.01.019.

[33] D. Ienco, R. Meo, Exploration and reduction of the feature space by hierarchical clustering, in: Proceedings of the 2008 SIAM International Conference on Data Mining, SIAM, 2008, pp. 577–587.

[34] J. H. Ward, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (1963) 236–244.

[35] K. Hopkins, G. Glass, Basic Statistics for the Behavioral Sciences, Prentice-Hall, Englewood Cliffs, N.J., 1978.

[36] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ, 1988.

[37] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, L.-P. Morency, Automatic audiovisual behavior descriptors for psychological disorder analysis, Image and Vision Computing 32 (2014) 648–658.

[38] S. Sorg, C. Vögele, N. Furka, A. Meyer, Perseverative thinking in depression and anxiety, Frontiers in Psychology 3 (2012). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00020. doi:10.3389/fpsyg.2012.00020.

[39] B. Li, D. Lima, Facial expression recognition via resnet-50, International Journal of Cognitive Computing in Engineering 2 (2021). doi:10.1016/j.ijcce.2021.02.002.

[40] V. Berisha, C. Krantsevich, P. R. Hahn, S. Hahn, G. Dasarathy, P. Turaga, J. Liss, Digital medicine and the curse of dimensionality, NPJ digital medicine 4 (2021) 153.