Multimodal prediction of obsessive-compulsive disorder, comorbid depression, and energy of deep brain stimulation

Saurabh Hinduja*, Ali Darzi*, Itir Onal Ertugrul, Nicole Provenza, Ron Gadot, Eric A Storch, Sameer A Sheth, Wayne K Goodman, and Jeffrey F Cohn#

Abstract—To develop reliable, valid, and efficient measures of obsessive-compulsive disorder (OCD) severity, comorbid depression severity, and total electrical energy delivered (TEED) by deep brain stimulation (DBS), we trained and compared random forests regression models in a clinical trial of participants receiving DBS for refractory OCD. Six participants were recorded during open-ended interviews at pre- and post-surgery baselines and then at 3-month intervals following DBS activation. Ground-truth severity was assessed by clinical interview and self-report. Visual and auditory modalities included facial action units, head and facial landmarks, speech behavior and content, and voice acoustics. Mixed-effects random forest regression with Shapley feature reduction strongly predicted severity of OCD, comorbid depression, and total electrical energy delivered by the DBS electrodes (intraclass correlation, ICC, = 0.83, 0.87, and 0.81, respectively. When random effects were omitted from the regression, predictive power decreased to moderate for severity of OCD and comorbid depression and remained comparable for total electrical energy delivered (ICC = 0.60, 0.68, and 0.83, respectively). Multimodal measures of behavior outperformed ones from single modalities. Feature selection achieved large decreases in features and corresponding increases in prediction. The approach could contribute to closed-loop DBS that would automatically titrate DBS based on affect measures.

Index Terms—Obsessive-Compulsive Disorder (OCD), Depression, Deep Brain Stimulation (DBS), Mixed-effects, multimodal machine learning, Shapley feature reduction

1 Introduction

Internalizing disorders (e.g., obsessive-compulsive disorder and depression) are characterized by anxiety, depressive, and somatic symptoms [1]. Advances in the development and provision of effective treatments for internalizing disorders depend on patient self-report and clinical interview. Self-report is limited by patients' reading ability, idiosyncratic use, inconsistent metric properties across scale dimensions, reactivity, and differences between clinicians' and patients' conceptualization of symptoms. Clinician interviews enable more consistent use, but are time-intensive, difficult to standardize across settings, inherently subjective, and susceptible to reactivity effects, rater drift, and bias. Neither self-report nor clinical interview have the granularity necessary to measure moment-to-moment response to intervention or enable brain-behavior quantification. To assess quantitative changes in treatment response, objective measures are needed.

Extant assessment methods fail to consider that internalizing disorders have marked observable influence on psychomotor functioning (e.g., agitation), expression of affect (reductions in positive affect and increases in negative), and interpersonal communication (lack of synchrony). Behavioral signal processing of audio and video recorded behavior has shown great potential to objectively measure symptoms of depression and to a lesser extent anxiety [2], [3], [4], [5], [6].

Further advances depend in part on four challenges. One is greater emphasis on severity rather than detection. While detection matters for screening purposes, to inform treatment and assess outcomes precise measurement of severity is what matters. For instance, percentage reduction in severity is a common measure of treatment response. Unless severity is measured, treatment response cannot be quantified.

Two is attention to internalizing disorders beyond depression. Depression is only one of many internalizing disorders that are cause for significant distress and disability. Internalizing disorders often are inter-related or comorbid as well. In the following work, we focus on participants with obsessive-compulsive disorder (OCD) that have comorbid symptoms of depression.

Three is inclusion of multimodal features. Multimodal communication encompasses facial expression, head and body motion, gaze, voice, speech behavior, and speech or language. Yet, as a recent scoping review [7] found, the vast majority of studies on depression include only one or two modalities. This is the case even though large numbers of studies have collected the necessary

* *Denotes equal contribution*
# *Corresponding author*

- *Saurabh Hinduja[1], Ali Darzi[2] and Jeffrey F Cohn[3] are with the Department of Psychology at the University of Pittsburgh, PA, 15213. Email:{[1]sah273, [2]ald260,[3]jeffcohn}@pitt.edu*
- *Itir Onal Ertugrul is with the Department of Information and Computing Sciences, Utrecht University, The Netherlands. Email: i.onalertugrul@uu.nl*
- *Nicole Provenza[4], Ron Gadot[5] and Sameer A Sheth[6] are with the Department of Neurosurgery at Baylor College of Medicine, TX, 77090. Email:{[4]nprovenz, [5]ron.gadot, [6]sasheth}@bcm.edu*
- *Eric A Storch[7] and Wayne K Goodman[8] are with the Menninger Department of Psychiatry and Behavioral Science at Baylor College of Medicine, TX, 77090. Email:{[7]storch, [8]wayne.goodman }@bcm.edu*

Manuscript received April 19, 2005; revised August 26, 2015.

audio-visual data [8]. Moreover, all but two of the studies that model multimodal data use a single corpus of distressed participants that lack clinical diagnosis and treatment. OCD and other internalizing disorders are much less studied. We focused on a clinical sample of participants diagnosed with refractory OCD and comorbid depression and in treatment.

And four, until recently [9], [10] investigators have typically treated multiple observations from the same persons as if they were independent [4], [11]. Failure to model the clustering of observations within persons ignores individual differences that may present confounds if not taken into account. When repeated observations within subjects are combined, trends may disappear or even reverse; an effect known as Simpson's paradox [12]. We compared multimodal random forest regression with and without *"random effects"*, where random effects account for the individual differences, to objectively measure change in severity within patients over the course of treatment for chronic, severe, obsessive-compulsive disorder.

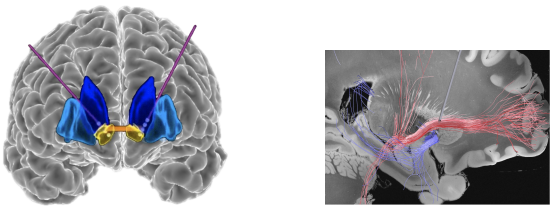
OCD is a persistent, oftentimes disabling condition that is characterized by obsessive thoughts and compulsions. Obsessions are repetitive and intrusive thoughts (e.g., contamination), images (violent scenes), or urges (e.g., to stab someone) that can be highly disturbing. Individuals with OCD attempt to ignore or suppress obsessions or to neutralize them with other thoughts or actions [13]. Compulsions are repetitive behaviors that an individual feels driven to perform in effort to reduce or avoid obsessions. Obsessions and compulsions are time-consuming (many hours per day), result in clinically significant impairment, and often are comorbid with depression, especially in more severe cases [14], [15].

Approximately 25-50% of OCD sufferers experience major depression [16], [17]. Core symptoms of each are conceptually different (i.e., obsessions and compulsions, versus dysphoria, anhedonia). Studies on the temporal nature of OCD and depression comorbidity suggest that in most (but not all) instances, obsessive-compulsive symptoms predate the depressive symptoms [18], [19].

Treatment-resistant OCD is defined as repeated failure to respond to front- or second-line treatments. Frontline treatments for OCD are cognitive-behavior therapy with exposure and response prevention (ERP), and serotonin reuptake inhibitors with or without clomipramine, a tricyclic antidepressant [20], [21]. Second-line treatments may include anti-psychotics [22]. About 25% of patients with OCD fail to respond to front- or second-line treatments or have difficulty with adherence or tolerance, respectively, and are considered treatment-resistant.

Deep brain stimulation (DBS) has shown promising results as a therapeutic intervention for patients with severe, treatment-resistant OCD. Participants were treated with DBS of or close to the ventral capsule/ventral striatum (VC/VS). The VC/VS is in a subcortical circuit involved in error detection, habit formation, and motivational processes [23], [24]. In studies by our group and others, DBS using implanted electrodes targeting nodes of this circuitry (Fig. 1 and 2) has proven highly effective in relieving treatment-resistant OCD. The most comprehensive and up to date review of DBS outcomes found that 66% of patients fully responded to treatment [25]. DBS also proved effective in treating depression; 50% of patients fully recovered from depression and another 16% partially recovered.

We measured change in severity of OCD and comorbid depression over the course of an 18-month clinical trial for treatment-resistant OCD. We tested the hypothesis that an unobtrusive AI-based system deployed in open-ended interviews can effectively yield biomarkers of OCD and comorbid depression severity as well as total electrical energy delivered (TEED) by the DBS electrodes. Participants undergoing DBS treatment for refractory OCD were recorded in open-ended interviews at regular intervals over the course of the trial. Modalities included facial expression, eye movement, head pose, voice acoustics and timing, and linguistic measures of speech. Because each participant was seen on a variable number of occasions, we used a mixed-effects random forest regression with feature reduction and cross-validation to control for individual differences and overfitting. To evaluate the effectiveness of random-effects, we used the same feature reduction and cross-validation with standard random forests. We seek to objectively measure response to treatment across the duration of clinical trials.



(a) (b)

Fig. 1: (a) Frontal view of an OCD patient's brain. Implanted DBS leads and their electrodes are shown with purple lines and white circles, respectively. The ventral striatum (target area) is in yellow. (b) Saggital view of the DBS electrodes in relation to the cortico-striatal-thalamo-cortical circuit that is implicated in OCD.

We first briefly review multimodal measures of affect related to internalizing disorders and novelties of the research and the research questions

1.1 Multimodal measures of affect

Extensive evidence in psychology and affective computing supports the view that affective communication is multimodal [2], [26], [27], [28], [29]. We briefly review literature relevant to both unimodal and multimodal communication of emotion and emotion disorders (aka internalizing disorders) such as depression, anxiety and OCD.

Visual features: The Facial Action Coding System (FACS) enables description of nearly all-possible visually discernible facial movement [30]. Movements for which the anatomic basis is known are referred to as Action units (AUs). Examples of AUs include AU 1 (medial strand of the frontalis, which raises the inner brow), AU 2 (lateral frontalis, which raises the outer brow), AU 6 (orbicularis oculi, which raises the cheeks, narrows the eye aperture and may cause "crows-feet" wrinkles at the lateral eye corners), and AU 12 (zygomatic major, which pulls the lip corners obliquely in smiling). While not without controversy, strong evidence suggests that specific combinations of actions are strongly related to specific emotions and intentions [31], [32], [33], [34], [35]. Automatic detection of AU occurrence and intensity and continuous measurement of some action descriptors has become possible [36], [37], [38], [39]. Velocity of automatically detected action units and head motion has been strongly related to emotional distress, depression, mania, and autism spectrum disorder [4], [29], [40], [41], [42], [43], [44].

Preliminary evidence suggests that facial AUs and head dynamics may differentiate between different levels of DBS stimulation [45] and predict OCD severity [46].

Acoustic features: Affective states strongly influence voice production [47], [48]. Change in subglottal pressure, transglottal airflow, and vocal fold vibration can be seen in acoustic features of affective speech. Additional features that have proven informative include vocal fundamental frequency (intonation and rhythm) [49], energy (volume or intensity) [50], utterance duration [51], and intra- and inter-speaker pause duration [50], [52]. Due to the effectiveness of acoustic and temporal features, they are frequently used in mental health studies: Anxiety [53], Distress Assessment [54], and depression and suicide [3]. Hence, acoustic and related temporal features are good candidates to assess DBS treatment in OCD patients.

Linguistic features: Linguistic features reveal sentiment and interests [55]. Prior to analysis, pre-processing is typically required, which includes localization of speakers' audio, speech recognition [56], and speech-to-text conversion [57]. To calculate linguistic features, several natural language processing techniques and models can be used. Notable examples include BERT [58], RoBERTa [59], PALM [60], cTAKES [61], and LIWC [55]. The instances of well-known linguistic features are syntax parsing using dependency trees, Chomsky transformational grammars, and statistical methods (e.g., word counting) [62]. Language-based deficits are common symptoms of psychiatric disorders [63]. Linguistic features are frequently used to detect depression and suicidal ideation [64], [65], [66], [67], [68], addiction [69], [70], [71], anxiety [72], and bipolar disorder [73].

Multimodal features: In social interaction, affective states are expressed multimodally. We considered facial actions, head motion, voice acoustics (e.g., vocal fundamental frequency), speech behavior (e.g, pause duration), and language (e.g., sentiment and word use, referred to below as text). Because these modalities may carry different messages, attention to a single modality can result in ambiguous or misleading results. To increase precision and accuracy, multimodal fusion can be performed. Feature-level fusion (or early fusion) [74], decision-level fusion (or late fusion), and hybrid-level fusion all may be useful. In early fusion, all features across modalities are placed together; and all or subsets are used to train a desired model. In decision-level fusion, separate modality-specific models may be developed and then fused using majority voting. Multimodal affective analysis can vary in the combination of modalities used to detect affective states. Several studies investigated how different modalities may complement each other to increase the performance of an ensemble model. For instance, combinations of acoustic-visual [54], acoustic-linguistic, or all three [4], [64] may be used. Most multimodal affective computing methods, using either early- or late fusion, typically outperform unimodal models.

1.2 Machine learning for internalizing disorders

Machine learning from behavioral features has been used widely to detect depression [75], [76], [3], [77], [78] and to some extent anxiety [54], post-traumatic stress disorder (PTSD) [79], [80], and suicidality [3]. Machine learning has been used less often to infer symptom severity, which is necessary to learn whether patients are improving or not with treatment.

Conventional machine learning approaches are based on designing and selecting hand-crafted features and training classifiers to detect disorders. Previous research has trained models including support vector machines (SVMs) [81], logistic regression [4], and decision trees [82] mostly with the aim of achieving high prediction performance. Deep learning approaches that automatically learn important features from the data often realize superior performance in detecting depressive [83] and manic episodes [84] compared to conventional approaches. However, a major drawback of deep learning based approaches is that large numbers of participants are required and features typically lack interpretability that is important for clinical science and treatment.

In clinical fields, a common goal is to develop a system that informs assessment, treatment, and mechanisms. To achieve a machine learning model with good performance in each of these areas, it is crucial to understand why a model has given a particular decision and which features are critical in evaluating the degree to which patients are improving or not. For that reason, recent works have revisited the use of hand-crafted features. They afford interpretable results and high predictive performance.

Shapley analysis has been especially informative in interpreting feature contributions to model performance [85]. Recent examples include mothers' depression in dyadic interactions with their adolescent offspring [29], mania prediction in bipolar disorder [64], and differentiation of apathy and depression in older adults [86].

As noted above, prior research in behavioral predictors of internalizing disorders often neglects to consider repeated assessments over time of the same individuals. When repeated assessments have been available, they have been treated as if they were independent [4]. When longitudinal assessments are available, attention to within-subject correlation is important to control for individual differences. For observations nested within individuals or treatment providers, mixed-effects models are needed. In mixed-effects models, each individual has their own, unique slope and intercept. Mixed-effects models are well known in behavioral statistics [87], [88] as multilevel models, but less so in machine learning. When multilevel structure is ignored, prediction may be impaired or confounded [89].

As an example, [10] compared standard random forests and mixed effects random forest for multiple observations within participants. Models with mixed effect performed better than those that lacked it. In another study [9], clustering of observations within treatment providers was the mixed-effect in GLMM and in random forests models. Both models provided good prediction but were not compared with models that failed to include a random effect. We hypothesized that with their ability to model the clustering of observations within participants, mixed-effects random forest regression improve performance compared with standard random forest regression.

We extend mixed-effects random forests three ways. First is to compare random forest models with and without random effects in predicting severity of OCD and comorbid depression. Second is to compare them in predicting total electrical energy delivered by DBS. Third is to evaluate the relative contribution of feature selection strategies for random forest models with and without random effects.

1.3 Novelties and Research Questions

This paper addresses point-wise severity of disorder and extends our preliminary work [46] in several ways:

1) Includes additional participants and observations over time.
2) Trains and compares random forest (RFs) models with and without random effects for data nested within participants. Participants were observed on 8 to 12 occasions over as much as 18 months.
3) Predicts OCD severity, comorbid depression severity, and energy delivered by DBS; and evaluates relative contributions of each set of features.
4) Uses Shapley analysis to reduce the number of features, optimize prediction, and afford interpretable parameters.

To our knowledge, this paper presents the first use of multimodal machine learning from affective behavior to assess severity of OCD, severity of comorbid depression, and level of DBS stimulation; and one of few to compare models with and without random effects for longitudinal data. In a clinical trial of 18 months duration with a single-subjects-with-replications design, each participant serves as their own control. Participants are a highly select group with treatment-resistant OCD that have been implanted with a deep brain stimulator The repeated measures design addresses the longitudinal demands of clinical research and treatment. A multidisciplinary team of psychiatrists, neurosurgeons, clinical psychologists, neuro-scientists, and engineers is actively involved in all phases of the study. Given the nature of the research, program officials from the U.S. National Institutes of Health (NIH) and Food and Drug Administration (FDA) are closely involved as well.

We address four research questions.

1- To what extent can severity of OCD and comorbid depression and DBS stimulation be predicted by visual (action units, head and face dynamics) and audio modalities (voice acoustics and speech behavior and speech content) independently? DBS stimulation is quantified as total electrical energy delivered (TEED)

2- Does random forest with random effects for participants better predict OCD, comorbid depression, and TEED than random forest without random effects?

3- Are multimodal models more predictive than unimodal models?

4- Which features within and across modalities are most predictive of symptom severity and TEED?

2 METHODS

2.1 Study Setup and Protocol

This study is from an ongoing clinical trial of DBS for treatment-resistant OCD. Inclusion criteria were: 1) Repeated failure to respond to evidence-based treatments (cognitive behavioral therapy and medication); and 2) severe symptoms as measured by a score greater than 27 on the Yale-Brown Obsession Compulsion Scale-I (YBOCS-I) (scale of 0-40). Consistent with the literature characterizing patients with treatment-resistant OCD [15], all participants evidenced some level of comorbid depression as measured by the Beck Depression Inventory (BDI-II) [90]. All diagnoses were made by an experienced clinician and confirmed by a second experienced clinician.

The first two participants were implanted with the Medtronic Activa PC+S DBS device; the other four were implanted with the Medtronic Summit RC+S Percept. With one exception, the three men and three women have completed at least 15 months of the study (Table I). A brief description of the study protocol follows.

Participants underwent a 1-month pre-implantation baseline evaluation followed by implantation of bilateral DBS electrodes in or near the VC/VS. A second baseline was observed prior to initial activation and programming of the DBS device. Patients then were seen for in-person or virtual visits monthly for open-loop programming of the DBS to optimize treatment. Each visit started with an open-ended interview with a clinician. The interviews were 3 to 8 minutes in duration and were followed by assessment of symptom severity using the YBOCS-II for OCD [91] and the BDI-II [90] for comorbid depression.

TABLE 1: Sessions available for analysis. Baselines 1 and 2 occurred before and after implantation of DBS electrodes, respectively, and prior to DBS activation.

Participant	S1	S2	S3	S4	S5	S6
Baseline1	✓	✓	✓	✓	✓	✓
Baseline2	✓	✓	✓	✓	✓	✓
3rd Month	✓	✓	✓	✓	✓	✓
6th Month	✓	✓	✓	✓	✓	✓
9th Month	✓	✓	✓	✓	✓	NA
12th Month	✓	✓	✓	✓	✓	NA
15th month	✓	✓	✓	✓	✓	NA
18th month	NA	✓	✓	✓	NA	NA

Interviews were conducted in a controlled environment free from any extraneous sources of noise. They were recorded using a GoPro camera and high-resolution microphone positioned about 10 to 15 degrees of frontal view of the patient. A separate GoPro camera recorded the interviewer. On the same or following day, the stimulation parameters were titrated as needed in what is referred to as a programming session. At approximately six months from study start, patients received Exposure and Response Prevention (ERP) therapy, a form of Cognitive Behavior Therapy, for two months. Over the course of the trial, we analyzed pre- and post-baseline interviews and interviews approximately every 3 months (Table 1). To consider possible within-session differences, each session was divided into two halves.

2.2 Total Electrical Energy Delivered (TEED)

DBS has three parameters. These are amplitude, pulse duration, and frequency. To combine these parameters as single metric, total electrical energy delivered per second, or power, was computed using the formula [92]:

$$TEED(W*1s) = I(A)^2.PW(sec).f(Hz).R(\Omega), \qquad (1)$$

where power is expressed in Watts, current in Amperes, pulse width in seconds, frequency in Hertz, and resistance in Ohms. Throughout the clinical trial, the stimulation frequency was held constant at 150.6 Hz. Since the purpose of the study was to predict variations in total electrical energy delivered, the constant term was omitted. Because measurement may be affected by transient fluctuations in battery output, pulse shape, or resistance, actual delivered energy may differ slightly from calculated values.

2.3 Study Setup and Protocol

Figure 2 depicts the analysis pipeline. Facial action units, head and face dynamics, and eye motion are extracted from video; acoustic and linguistic features are extracted from audio. Using individual sets of extracted features, models are trained to predict OCD severity, comorbid depression severity, and TEED. SHAP analysis is used to evaluate the most informative unimodal features. We then aggregate the top-k features from each set and train a multimodal model. Finally, with a SHAP analysis on multimodal features,

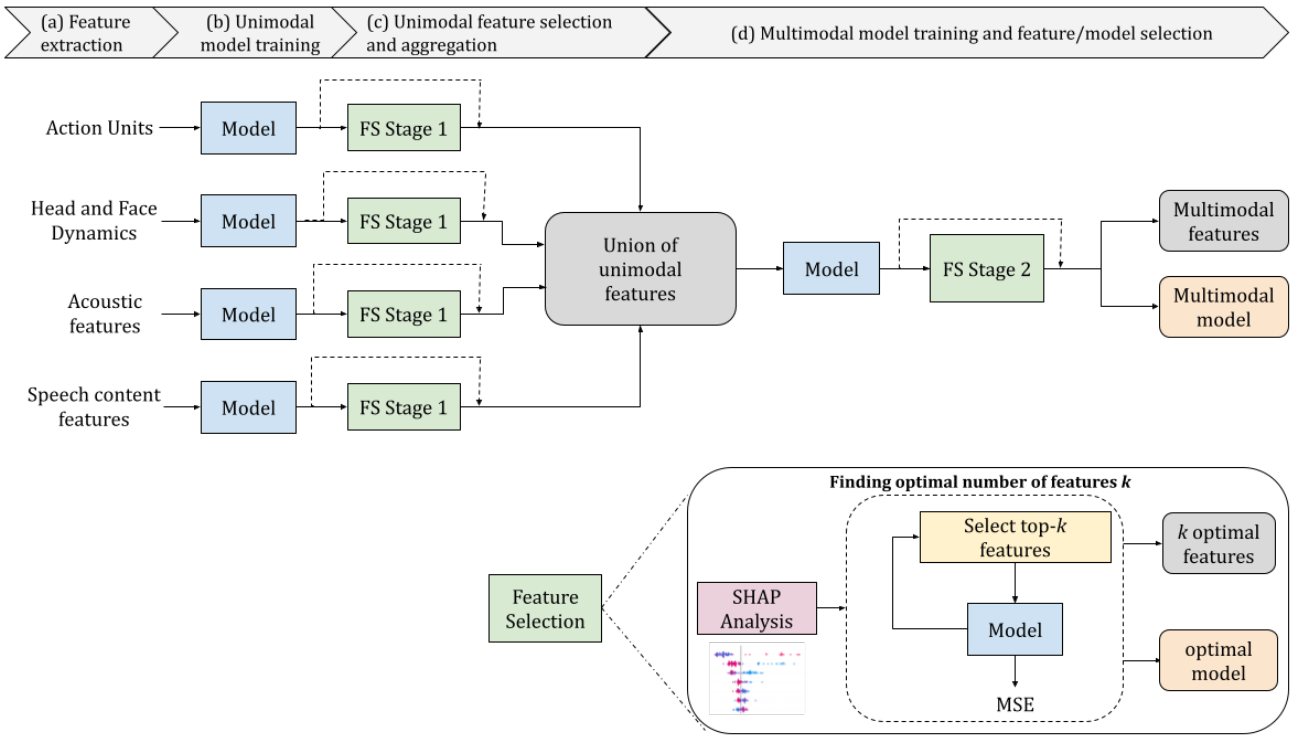


Fig. 2: Pipeline. (a) Action units, head and face dynamics, acoustic features, and speech content features are extracted. (b) Models (random forest regressions with and without sessions as mixed effects) are trained using each of the four feature sets separately. (c) Feature selection Stage-1 is completed. The most informative k features are selected from each of the four sets of features. Feature selection consists of SHAP analysis to identify informative features and rank them by their Shapley values; choose k optimal features for each feature set; and aggregate them into a single feature vector. (d) Models are trained with a combined multimodal set of optimal features. (e) Optimal multimodal features and the corresponding model are found with a multimodal feature selection step at Stage 2.

we identify the most-informative k multimodal features and the corresponding model. Because subject identity was important for mixed effects modeling, we used leave-one-session-out cross validation for all our experiments.

2.4 Multimodal Features

We extract four sets of features namely action units, head and face dynamics, acoustic features, and linguistic features.

Action units: Faces in the video are tracked and normalized using the Zface module of AFAR [36]. ZFace [93] is a real-time face alignment software that accomplishes dense 3D registration from 2D videos and images without requiring person-specific training. Faces are normalized to have an interocular distance (iod) of 80 pixels. AU detector module of AFAR is used to detect facial action units (AUs) in the normalized faces.

The version of AFAR, used in this study, was trained on the EB+ dataset (an expanded version of BP4D+ [94]), in which participants interact with an experimenter in a variety of emotion related tasks. Reliability of AFAR in EB+ was tested using k-fold cross validation. Average free-margin kappa was 0.75 and AUC 0.73 [37]. Cross-domain generalization was assessed by testing AFAR in Sayette GFT [95]. Average free-margin kappa was 0.49 and AUC 0.66, which represent moderate cross-domain generalizability. Because test results in GFT were likely attenuated by the larger head motion and lower video resolution in GFT, these comparisons provide a conservative estimate of the cross-domain generalizability in the current study. EB+ and the clinical trial were more alike than EB+ and GFT. EB+ and the clinical trial both used higher resolution video and were more similar in their more limited head motion.

AFAR was used to asses intensity of 6 facial AUs: AU1 (inner brow raiser), AU6 (cheek raiser), AU10 (upper lip raiser), AU12 (lip corner puller), AU14 (dimpler), AU17 (chin raiser). AU 1+2 is typically seen in surprise and affective engagement. An additional feature was average intensity of AU 6+12, which comprises the Duchenne smile, a marker of positive affect. For each of these, we extract time-series featues using tsfresh [96]. TsFresh outputs 794 time series characteristics for each feature for a total of 5,558 (7×794) features. In case of tracking failure or an AU feature fails to vary throughout the video, all related TsFresh features are set to 0. While the number of features is initially large, the analysis plan greatly reduces the number in a number of unimodal and multimodal steps.

Head and face dynamics: Head dynamics is defined using the time series of the 3 degrees of freedom of out-of-plane rigid head movement, which correspond to head nods (i.e., pitch), head turns (i.e., yaw), and lateral head inclinations (i.e., roll). Face dynamics is defined as time series of per frame eye and mouth openings. Eye opening is calculated using the Eye Aspect Ratio (EAR) [97], which is a normalized measure that divides

the distance between landmarks on the upper and lower eyelids to the distance between inner and outer eye corners. Average of left and right EAR is used. Similarly, mouth opening is calculated using the Mouth Aspect Ratio (MAR), which is a normalized measure that divides the distance between landmarks on the upper and lower lips to the distance between left and right mouth corners. After head (pitch, yaw, and roll) and face (EAR, MAR) dynamics are calculated, time series characteristics are extracted using TsFresh [96], yielding a total of 3,970 features.

Acoustic features: Audio for each speaker is localized and transcribed using TranscribeMe [98]. Audio and text are aligned using the Montreal-Forced-Aligner [99]. The openSMILE [100] toolkit and Collaborative Voice Analysis Repository (COVAREP) [101] are used to extract acoustic features. For openSMILE, we use the Geneva minimalistic acoustic parameter set (eGeMAPS [102]), which is a subset of audio features chosen for their ability to represent affective physiological changes in voice production. eGeMAPS contains 62 features: arithmetic mean and coefficient of variation of 18 low-level descriptors (LLD), 8 functionals applied to loudness and pitch LLD, and 6 temporal features. COVAREP provides 72 low-level speech acoustic features, which are derived from the speech signal, that include pitch, energy, spectral envelope, loudness, voice quality and other characteristics. Both eGeMAPs and COVAREP have been used extensively in the analysis of psychological disorders [103], [104], [105] and affect recognition [106], [107]

LIWC features: We use Linguistic Inquiry and Word Count (LIWC) [108], [109], which is a text analysis tool that determines the percentage of words in a text that fall into one or more linguistic, psychological, and topical categories. We extract 92 features from the verbal content of each interview. Approximately 93% of the words used in each interview were present in the LIWC dictionary and analyzed. We drop the coverage variable (referred to as "Dic" in LIWC) and normalize the word count variable with the duration of interaction.

2.5 Unimodal model training

When measures cluster within persons, as in a longitudinal study, mixed effects models are used in statistics and econometrics [110], [111]. In addition to the fixed effect terms, they include random effect parameters, which change the model's assumptions to accommodate heterogeneous data with many sources of random variability (e.g., both intra- and inter-individual). As a result, mixed effects methods allow for more accurate statistical inferences about the factors that connect with observed variance.

Motivated by two previous studies that used mixed effects models to infer depression severity in a longitudinal design [10], [46], we use MERF to infer OCD severity, related depression severity, and TEED.

MERF [111] is defined as:

$$Y_{ij} = f(X_{ij}) + b_i + \varepsilon_i \tag{2}$$

where: $i = 1, ..., m$ are the *clusters* (participants) each with n_i observations ($j = 1, ..., n_i$); X_{ij} is the input feature matrix; $f(X_{ij})$ is the *fixed effects random forest*; b_i is the *random effect* parameter; ε_{ij} is the measurement error; and Y_{ij} is the regression target variable. In our unimodal experiments, the fixed effect parameters are the features derived from individual modalities and the random effect parameter is the participant ID. The model is trained using expectation maximisation (EM) with a generalised log-likelihood (GLL) function to monitor convergence. For each cross- training / testing, the mixed effects random forests were trained for 50 iterations. To compare the effectiveness of mixed-effects random forests, we also trained standard random forests (RFs) using the same protocols. For standard random forests, we performed a grid search to find the best hyperparameters. For consistency of results across experiments and folds, the hyperparameters were kept constant (number of estimators at 250, criterion as poisson, max depth as none).

2.6 Unimodal feature selection

Shapley values [112] were introduced in game theory to gauge each player's participation in cooperative games. The machine learning and explainable AI communities recently have shown interest in Shapley. Shapley value for the j^{th} feature is defined as the weighted average of differences in predictions in the presence of the j^{th} feature and when it is marginalized, given the i^{th} data instance with m features represented by ${X_i}^m$. Marginalization is accomplished by leveraging predictions from several feature subsets. Calculating Shapley value is computationally expensive due to the laborious marginalization procedure. However, the SHAP (SHapley Additive exPlanations) framework can be used to estimate Shapley values [113]. Shapley value φ_j of feature j can be computed as:

$$\varphi_j(v) = \sum_{S \subseteq \{1,2,...,m\} \setminus \{j\}} \frac{|S|!\,(m-|S|-1)!}{m!} (v(S \cup \{j\}) - v(S)),$$

where v is the model, m is the total number of features and S is a subset of features.

We use kernel-based LIME, which combines Shapley values with Local Interpretable Model-agnostic Explanations (LIME) [114]. LIME has been widely used to interpret model decisions in the explainable AI field. While LIME provides local correctness, the SHAP framework improves upon that by ensuring feature consistency and robustness to missing features. Missing features have no impact on the contribution of features of interest. The SHAP analysis is done independently of the model training processes. The model itself (not the output of the model) is given as input for the analysis.

We use SHAP values to rank characteristics in terms of their relative contribution to prediction performance. We then choose the top-k features, where the optimal value of k is found iteratively based on the mean square error of the model trained with top-k features. Optimal k may differ for each set of features (e.g. action units, head and face dynamics, acoustic features, and linguistic features). We refer to this stage of feature selection as *"Feature Selection Stage 1"*. Since feature selection is done independent of model training, a new model is trained for each distinct set of features. This feature selection methodology is consistent with [103], [115], [116], and [29]. Optimal features for all individual modalities are concatenated to obtain union of unimodal optimal features (see Figure 2c)

2.7 Multimodal model training and feature/model selection

Following unimodal modeling, we train the models using multimodal features. Multimodal features comprise the combined top-k features selected from action units, head and face dynamics,

acoustics, and linguistics. By training models with selected multimodal features, we seek to reveal the relative contribution of each modality to performance. We use combined features as the fixed effects and participant ID as the random effect parameter.

We use SHAP values to rank the multimodal features based on their relative contribution to the prediction performance. Similarly to unimodal feature selection, we choose top-k multimodal features and optimize the value of k using an iterative approach. We refer to this stage of feature selection as *"Feature Selection Stage 2"*

2.8 Model training and evaluation

We trained separate models to predict OCD severity, comorbid depression severity, and TEED. For the multimodal model and each of the unimodal models, we optimized the number of features k in the set $k \in \{6, 11, 16, ...46\}$.

To evaluate relative performance of the models we used the following performance metrics:

1) *Mean Absolute Error (MAE)* is one of the most commonly used performance metric for continuous labels. It is defined as the sum of the absolute errors divided by the number of observations.
$$MAE = \frac{\sum_{i=1}^{D} |x_i - y_i|}{D} \tag{3}$$
where D is the number of observations, x_i is the ground truth score, y_i is the predicted score.
2) *Root Mean Squared Error (RMSE)* is the root of the mean of the square of the errors. RMSE score can never be zero. It is a frequently used metric for continuous observations.
$$RMSE = \sqrt{\frac{\sum_{i=1}^{D} (x_i - y_i)^2}{D}} \tag{4}$$
3) *R Square (R^2)* is also known as the coefficient of determination. It is always in a range (0,1).
$$R^2 = 1 - \sqrt{\frac{\sum_{i=i}^{D} (x_i - y_i)}{\sum_{i=i}^{D} (x_i - \bar{y})}} \tag{5}$$
where $\bar{y}$ is mean ground truth score.
4) *Intraclass Correlation (ICC)* is commonly used to determine the correlation between raters. In our case, "raters" are represented by ground-truth and predicted scores. ICC may be computed for agreement or for consistency. We used ICC(2,1) for consistency. To evaluate the predictive power of the model, we performed F-test for ICC scores. Because of the large number of tests (n=24 for each predicted measure) p was set as less than 0.001 .
5) *Normalized mean absolute error* is the ratio of MAE to the range of measure (ROM), which is the difference between the possible maximum and minimum values of the measure. It is calculated as:
$$Norm\_MAE = \frac{MAE}{ROM} \tag{6}$$
As the ranges of the OCD, comorbid depression, and TEED measures differ, directly comparing MAEs obtained with each of them would not be meaningful. By dividing the MAEs by the range of each measure, we obtain a normalized measure that is comparable across all of them.
6) *Contribution of features* is used to find the importance of a modality in the prediction of the model. It is based on SHAP values. We define contribution for a particular modality as:
$$Contribution = \frac{\sum_{j=1}^{F} \sum_{i=1}^{D} SM_{ij}}{\sum_{k=1}^{M} \sum_{j=1}^{F} \sum_{i=1}^{D} S_{ijk}} \times 100 \tag{7}$$
where SM is SHAP value for a particular modality, S are all SHAP values, F is number of features in a modality, M is number of modalities. It is the ratio of sum of the all the SHAP values for a particular modality by the sum of all SHAP values across all modalities.
7) *Contribution of feature selection* is evaluated using the Wilcoxon Test [117]. The Wilcoxon Test is a non-parametric statistical test used to compare paired samples and assess whether there is a significant difference in their distributions. For each of the predicted measures (i.e., OCD, comorbid depression, and TEED), 12 models used feature selection and 12 models did not use feature selection. We tested the hypothesis feature selection would result in a higher ICC score than models that lacked feature selection. For each of the three measures, we used a *p-value* of less than 0.05. The Wilcoxon Test statistic is calculated as follows:
$$W = \sum_{i=1}^{n} \text{sgn}(X_i - Y_i) \cdot \min(|X_i - Y_i|, 0.5) \tag{8}$$
where n is the number of paired observations, X_i and Y_i are the paired observations from the two groups, and $\text{sgn}(\cdot)$ is the sign function. The resulting W statistic is then compared to critical values to determine the statistical significance of the observed differences.
8) *Contribution of random effects*. The contribution of random effects was similarly evaluated using the Wilcoxon Test [117]. We tested the hypothesis that models with random effects would result in a higher ICC score than models that lacked random effects with *p-value* of 0.05.

3 RESULTS

In Section 3.1, we report MERF and RF results for OCD severity, comorbid depression severity, and total energy delivered (i.e., YBOCS II, BDI II, and TEED, respectively). These include the test statistics (e.g., ICC) for each model and modality. In Section 3.2, we present the most important SHAP identified features within each modality.

3.1 Prediction results

OCD severity: The top of Table 2 shows the performance of each of the unimodal models; MERF and RF, with and without feature selection stage-1.

Among the unimodal models trained without Feature Selection Stage-1, the MERF model trained with acoustic features gave the best performance on each of the performance metrics. ICC for MERF trained with acoustic features is 0.76 and ICCs for other unimodal models ranged from 0.25 - 0.48. For RFs (i.e., random forests that lack mixed-effects), acoustic features performed the best with ICC of 0.51 and other fixed-effects unimodal RFs ranged from 0.03 - 0.21.

Feature selection stage-1 improved performance for each of the unimodal models. MERF model with acoustic features was the best among them and required only six features.

TABLE 2: Prediction of OCD severity using unimodal and multimodal features. OCD was measured using the YBOCS-II. *Note: The best performing model is indicated in Bold and second best performing model is indicated in Italics. The significant ICC scores are indicated with * for p-value less than 0.001.*

	Modalities	Models	Feature Selection Step 1	Number of Features	MAE	RMSE	R^2	ICC
Unimodal	**Acoustic**	Random Forest	No	134	7.84	9.27	0.37	0.51*
			Yes	11	6.59	8	0.53	0.69*
		Mixed Effects RF	No	134	5.93	7.29	0.61	0.76*
			Yes	**6**	**5.05**	**6.22**	**0.72**	**0.84***
	LIWC	Random Forest	No	92	9.62	11.18	0.08	0.21
			Yes	11	8.42	9.65	0.3	0.42*
		Mixed Effects RF	No	92	8.67	10.26	0.2	0.45*
			Yes	11	7.85	9.58	0.32	0.57*
	AUs	Random Forest	No	5,558	10.54	12.24	0.11	0.03
			Yes	26	7.41	9.69	0.35	0.47*
		Mixed Effects RF	No	5,558	8.67	10.29	0.22	0.46*
			Yes	36	6.7	8.1	0.5	0.7*
	Head & Face	Random Forest	No	3,970	10.01	11.52	0.02	0.1
			Yes	21	7.51	9.17	0.38	0.51*
		Mixed Effects RF	No	3,970	8.46	10.25	0.23	0.48*
			Yes	16	5.86	7.67	0.57	0.74*
Multimodal	**Union**	Random Forest	No	9,754	10.05	11.48	0.02	0.11
			Yes	69	7.78	9.16	0.38	0.48*
		Mixed Effects RF	No	9,754	8.57	10.3	0.21	0.48*
			Yes	69	7.65	9.82	0.67	0.81*
	Union with Feature Selection Step 2	Random Forest	No	16	7.16	8.54	0.46	0.59*
			Yes	16	6.87	8.23	0.5	0.62*
		Mixed Effects RF	No	16	5.08	6.32	0.7	0.82*
			Yes	*6*	*5.2*	*6.47*	*0.7*	*0.83**

TABLE 3: Prediction of comorbid depression severity using unimodal and multimodal features. Comorbid depression was measured using the BDI-II. *Note: The best performing model is indicated in Bold and second best performing model is indicated in Italics. The significant ICC scores are indicated with * for p-value less than 0.001.*

	Modalities	Models	Feature Selection Step 1	Number of Features	MAE	RMSE	R^2	ICC
Unimodal	**Acoustic**	Random Forest	No	134	12.94	15.27	0.19	0.36
			Yes	6	10.34	13.19	0.4	0.57*
		Mixed Effects RF	No	134	8.14	10.29	0.64	0.8*
			Yes	26	7.9	10.29	0.64	0.8*
	LIWC	Random Forest	No	92	13.74	16.03	0.11	0.23
			Yes	21	12.34	14.59	0.26	0.4*
		Mixed Effects RF	No	92	8.83	10.4	0.18	0.44*
			Yes	21	11.05	11.05	0.39	0.62*
	AUs	Random Forest	No	5,558	12.59	15.59	0.15	0.3
			Yes	16	10.14	12.78	0.44	0.61*
		Mixed Effects RF	No	5,558	11.65	13.88	0.33	0.57*
			Yes	10	8.1	10	0.6	0.8*
	Head & Face	Random Forest	No	3,970	13.24	15.28	0.16	0.29
			Yes	11	9.83	12.56	0.46	0.62*
		Mixed Effects RF	No	3,970	11.83	14.05	0.32	0.56*
			Yes	16	8.67	10.76	0.6	0.75*
Multimodal	**Union**	Random Forest	No	9,754	13.46	16.34	0.08	0.22
			Yes	50	9.87	11.97	0.51	0.62*
		Mixed Effects RF	No	9,754	9.66	12.17	0.49	0.7*
			Yes	81	6.55	8.59	0.76	0.86*
	Union with Feature Selection Step 2	Random Forest	No	11	11.58	11.58	0.57	0.67*
			Yes	16	9.12	10.97	0.58	0.7*
		Mixed Effects RF	*No*	*21*	*6.53*	*8.57*	*0.75*	*0.86**
			Yes	**11**	**6.3**	**8.28**	**0.76**	**0.87***

TABLE 4: Prediction of Total Electrical Energy Delivered (TEED) using unimodal and multimodal features. *Note: The best performing model is indicated in Bold and second best performing model is indicated in Italics. The significant ICC scores are indicated with * for p-value less than 0.001.*

	Modalities	Models	Feature Selection Step 1	Number of Features	MAE	RMSE	R^2	ICC
Unimodal	**Acoustic**	Random Forest	No	134	3.07	3.74	0.44	0.63*
			Yes	16	2.71	3.33	5.56	0.72*
		Mixed Effects RF	No	134	3.22	3.88	0.34	0.31
			Yes	20	2.67	3.28	0.53	0.71*
	LIWC	Random Forest	No	92	4.18	5.18	0.08	0.05
			Yes	6	3.36	4.36	0.24	0.42*
		Mixed Effects RF	No	92	3.71	4.67	0.02	0.27
			Yes	6	3.34	4.35	0.24	0.46*
	AUs	Random Forest	No	5,558	4.11	5.14	0.06	0.11
			Yes	6	2.9	3.79	0.42	0.62*
		Mixed Effects RF	No	5,558	4	4.7	0	0.3
			Yes	21	3.05	4	0.36	0.54*
	Head & Face	Random Forest	No	3,970	3.47	4.32	0.25	0.36
			Yes	16	2.52	9.71	0.61	0.75*
		Mixed Effects RF	No	3,970	3.66	4.74	0.02	0.24*
			Yes	10	2.78	3.54	0.45	0.62*
Multimodal	**Union**	Random Forest	No	9,754	3.02	3.52	0.5	0.66*
			Yes	44	2.48	3.03	0.63	0.76*
		Mixed Effects RF	No	9,754	3.25	3.91	0.39	0.59*
			Yes	56	2.7	3.3	0.6	0.7*
	Union with Feature Selection Step 2	Random Forest	No	*21*	*2.17*	*2.65*	*0.72*	*0.82**
			Yes	**10**	**2.1**	**2.66**	**0.72**	**0.83***
		Mixed Effects RF	No	6	2.42	2.86	0.67	0.80*
			Yes	16	2.33	2.79	0.69	0.81*

By comparing the union of features and unimodal models with feature selection stage-1, we can evaluate whether union of features improved performance. For acoustic features, multimodal model afforded no advantage. The acoustic MERF model with feature selection stage-1 was equal to or outperformed the multimodal model on all four performance metrics. For other unimodal models with feature selection stage-1, the difference between unimodal and multimodal were mixed.

By comparing "Union" and "Union with Feature Selection Stage-2" we can evaluate whether further SHAP reduction is valuable.

The MERF Model with both feature selection stage-1 and stage-2 achieved the highest performance with an ICC of 0.83 and a large reduction in the number of features. Both feature selection stages in the multimodal model optimized prediction of OCD severity.

The ICC scores of MERF were significantly higher than those for RF by Wilcoxon Test. This finding suggest that random effects significantly contributed towards the prediction.

We compared the ICC score of models with and without feature selection stage-1; the resulting p-value was significant, indicating that SHAP reduction boosted model performance.

To evaluate MERF's prediction accuracy for OCD severity in individual participants, we computed participant-level ICCs. For 4 of 6 participants (S1, S3, S4, and S5), ICCs were statistically significant ($p < 0.01$) within the range ICC = 0.57 to ICC = 0.79.

Comorbid depression severity: Correlation between severity of OCD and comorbid depression ranged from moderate (r = 0.53) to strong (r = 0.90). Because the measures were not highly correlated for all participants and because predictive features, inclusion or not of random effects and feature selection could differ for the two measures, we report results separately for severity of comorbid depression.

Table 3 shows the corresponding results for comorbid depression (BDI II). As for OCD, the unimodal model for MERF with acoustics again was the best performing with an ICC of 0.80. ICCs for the unimodal models were much lower and similar trends were found for other test metrics.

Feature selection stage-1, again, statistically improved performance for both MERF and RF models. Unimodal MERF with feature selection stage-1 using acoustics outperformed other unimodal models. The number of features required for acoustic MERF, however was larger than that for OCD prediction (Table 2).

As in Table 2, by comparing "Union" and "Union with Feature Selection Stage-2", we can evaluate whether feature selection stage-2 improved performance relative to feature selection stage-1. In contrast to the results for OCD, feature selection stage-2 for comorbid depression outperformed that for acoustics and other modalities.

Wilcoxon test, for contribution of random effects showed a statistical improvement in performance with a p-value of less than 0.05.

To evaluate MERF's prediction accuracy for severity of comorbid depression in individual participants, we computed participant-level ICCs. For 4 of 6 participants (S2, S3, S4, and S5), ICCs were statistically significant ($p < 0.01$) within the range ICC = 0.52 to ICC = 0.74.

Total Electrical Energy Delivered (TEED): Table 4 shows the corresponding performance for TEED. Acoustic models achieved

the highest prediction. RF performed better than MERF. Random effects made no significant contribution (p-value 0.876).

Similar to OCD and comorbid depression, feature selection stage-1 statistically improved the performance for both MERF and RF. For TEED, RF using Union with feature selection stage-1 and stage-2 performed the best. The next best model was RF using Union with only feature selection stage-2.

To evaluate MERF's prediction accuracy for TEED in individual participants, we computed participant-level statistics. For 4 of 6 participants (S3, S4, S5, and S6), ICCs were statistically significant ($p < 0.01$) within the range ICC = 0.73 to ICC = 0.84.

Performance comparison across labels: To further compare performance across OCD, comorbid depression, and TEED, we compute the normalized MAE values for each individual set of features and their combination as shown in Figure 3. For MAE, lower scores are better. For individual modalities, acoustic features yielded the lowest MAE for OCD, comorbid depression, and TEED. Linguistic features yielded the highest. Multimodal models with SHAP reduction yielded the smallest MAEs for comorbid depression and TEED although not OCD.

Prediction of OCD severity and comorbid depression severity (Figure 3a and Figure 3b) are similar for all feature types except for acoustic features, which perform better. TEED prediction performances given in Figure 3c are much lower compared to other two for all feature sets.

3.2 Relative contribution of features within modalities for the best-performing multimodal model

For OCD severity, comorbid depression severity, and TEED, Table 5 presents the top-*k* multimodal features in predicting their respective values. Red denotes that an increase in the value of the feature leads to an increase in the predicted value. Blue indicates that an increase in the value of the feature leads to a decrease in the predicted value.

For severity of OCD and comorbid depression, most of the top-*k* multimodal features were acoustic. Of these, MFCCs were especially informative. MFCCs are a non-linear transformation of the audio signal that approximate human perception. MFCCs 4, 6, and 8 were among the most informative for severity of both OCD and comorbid depression. While individual MFCCs are difficult to interpret, coefficients from the lower half of the range suggest psychomotor retardation, vocal track constriction, and reduced energy [2], [86]. Harmonic and phase distortion coefficients were negatively correlated with comorbid depression severity.

Head and face dynamics were related to comorbid depression and TEED but not OCD. With increases in TEED, mouth, eye, and brow movement became more frequent. With increases in comorbid depression they decreased.

Brow raises (indexed by AU 1) were negatively correlated with OCD severity. Brow raises are an affiliative display indicative of interest. Linguistic features did not contribute to OCD, comorbid depression or TEED.

3.3 Individual differences among participants

To visualize individual differences, we plotted the predicted values by ground truth of the best multimodal performing model for OCD, comorbid depression, and TEED (Figure 4a, Figure 4b, and Figure 4c). For OCD the best performing multimodal model was MERF with feature selections. With exception of S6, the slopes for OCD were closely related in intercepts and slopes across participants. For S6, a factor may have the smaller number of observations and attenuated variability of OCD for them.

For comorbid depression the best performing multimodal model was MERF with feature selection. In the predictions, there was more variability across participants in slopes and intercepts. Attenuated variability may have contributed to this finding especially in S6. These findings support the importance of mixed-effects modeling for longitudinal machine learning.

For TEED the best performing model was RF with feature selection. The slope between the predicted and ground truth values for TEED was always positive with some variability in the slopes of across participants. As random effects were not considered for TEED prediction, number of sessions for the subject was not a contributing factor and the slope of S6 was also positive. The slope had high variability.

4 Discussion

Clinician ratings of OCD and comorbid depression were inferred with high reliability from multimodal RF models. For OCD severity, the ICC for the SHAP-reduced ICC model was 0.83; for depression, the corresponding ICC was 0.87. These ICCs rival the interrater reliabilities of trained clinicians. This strong performance was enabled by including random effects in the models. Random effects account for individual differences in participants. When random effects were omitted, prediction decreased (ICC = 0.60 to 0.68). This finding suggests that individual differences in participants are important to consider for longitudinal measures.

Total energy delivered by the DBS electrodes was also strongly predicted by multimodal RF. Unlike for symptom ratings, however, inclusion of random effects provided little or no advantage. Across individual modalities and in multimodal models, RF performed comparable to or marginally better than MERF. The ICC for the SHAP-reduced model was 0.83. This finding suggests that individual differences in participants are less consequential for TEED.

For OCD, comorbid depression and TEED, feature reduction using SHAP consistently outperformed models that omitted feature selection. In most cases the advantage of feature selection dramatically improved model performance. These findings are consistent with recent research. Alghowinem and colleagues [118] found that feature selection improved model performance in independent corpora from three countries; Bilalpur found benefit from feature reduction in mothers with depression [29]. Similar gains have been realized in affect recognition [119]. Given the consistency of the findings between corpora and contexts, we would anticipate similar benefits from feature selection in other clinical datasets. This hypothesis requires testing. A related question is what specific features are robust to differences in data sources.

With further research, user-in-the-loop systems could be an important outcome. A critical problem in clinical trials research is lack of consistency among raters on endpoint measures such as the HAM-D and YBOCS interviews. Even when raters are well trained and interrater reliability is monitored carefully, agreement tends to drift over time especially across different data collection sites. Inadequate reliability leads many clinical trials to be underpowered. Because aggregating independent measurements is an effective means to increase reliability of measurement [120], aggregating computational and human ratings could boost the effective reliability of symptom ratings in clinical trials. Another possible user-in-the-loop application would be to flag ratings that

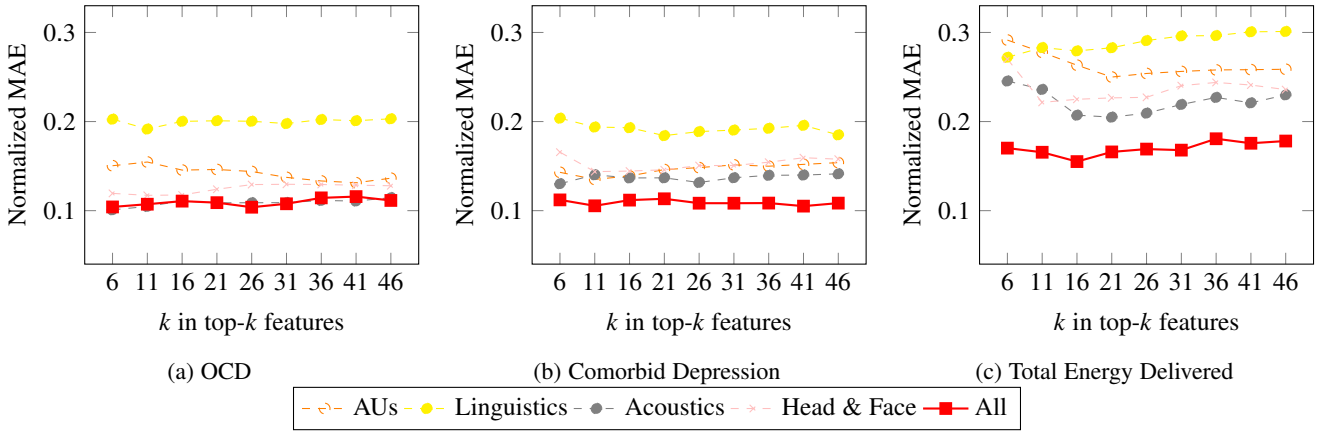


Fig. 3: Cross-validation Normalized MAE performance of top-k features derived from SHAP analysis

TABLE 5: SHAP ordering of the top-k features across all modalities in predicting symptom severity of OCD, symptom severity of comorbid depression, and total electrical energy delivered (TEED) by the DBS electrodes. *Note: The color indicates the sign of the correlation; red for positive correlation and blue for negative correlation.*

Modality	Feature	OCD (MERF)	Comorbid Depression (MERF)	TEED (RF)
Acoustic	MFCC 4	1	4	3
	Loudness	2	-	-
	HMPDM 13	3	1	-
	MFCC 6	5	8	-
	MFCC 8	6	3	-
	MFCC 18	-	2	1
	HMPDM 8	-	6	-
	MFCC 24	-	7	10
	MFCC 20	-	-	8
	MFCC 13	-	11	-
	HMPDM 10	-	-	7
Head and Face Dynamics	Yaw autoregressive likelihood	-	-	4
	Mouth approximate entropy	-	-	9
	Mouth mean of change in quantiles	-	-	5
	Mouth variance of change in quantiles	-	-	6
	Eyes permutation entropy	-	5	-
	Eye change in quantile of mean	-	10	-
	Eye velocity of change in opening	-	9	-
	Mouth longest strike below mean	-	-	2
Action Units	AU1 Benford correlation	4	-	-

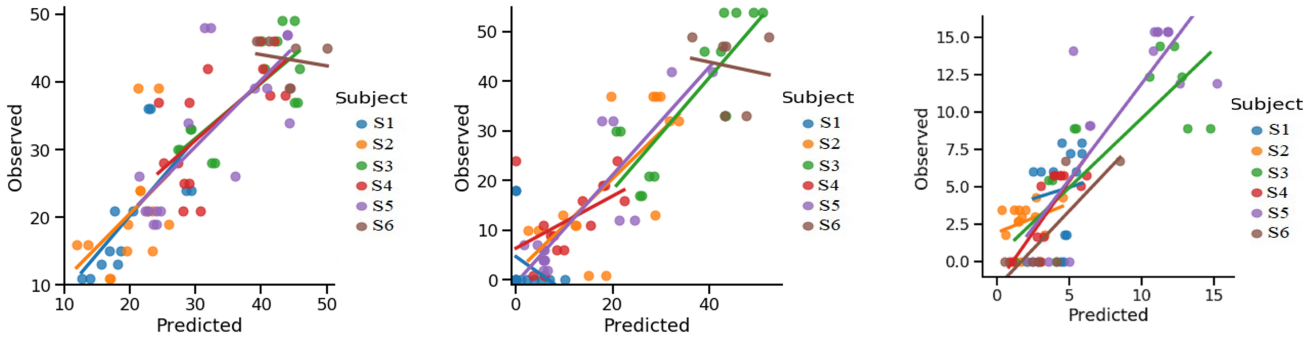
are too discrepant from model predictions, which would enable review and timely re-interviewing as warranted.

In treatment settings, computational measurement of severity could provide automated measurement of symptom severity over the course of treatment to answer the question of whether patients are improving or not and to inform therapy. Until now, subjective self-report measures and clinical judgment have been the only means to gauge treatment response in therapy. In addition to their limitations, noted above, completion rates for self-report measures are known to decrease when repeated over time [121].

In the neuroscience community, there is increasing interest in brain-behavior quantification and synchronization [122]. That is, how changes in neural activity relate to synchronous changes in behavior. For relatively brief interviews that we tested, we found strong correlation between total energy delivered by the DBS electrodes in or near the VC/VS and session-level multimodal behavior. An exciting next step is to test whether synchronization of neural stimulation and multimodal behavior occurs across briefer intervals. Initial work with [123] is exploring this question.

The current state of the art in DBS for treatment refractory OCD is open-loop programming. That is, patients return to the clinical setting at frequent intervals to evaluate recent symptoms and adjust DBS parameters as needed. In these sessions, clinical interviews and observations inform determination of DBS parameters through trial and error. Judgments are subjective and vary within and between clinicians and over time. A multimodal regressor could greatly reduce or eliminate subjective judgment and enable more accurate titration of the DBS.

Because exposure to OCD triggers and the severe symptoms they elicit can vary within and across days, more frequent evaluation and adjustment of DBS parameters than is possible in periodic office visits would be beneficial. In DBS-treatment for essential tremor, effective closed-loop programming has been achieved. The same is a current goal of research in DBS treatment for refractory



(a) OCD (MERF) (b) Comorbid Depression (MERF) (c) TEED (RF)

Fig. 4: Observed and Predicted values for the best performing model

OCD. The current findings suggest that multimodal behavior acquired via audio and video could be an effective component of a self-titrating DBS system. It also could be effective in detecting hypomania or mania, which are side effects of DBS, and automatically down-regulating DBS energy to reduce or eliminate this unwanted and potentially dangerous side effect.

An unexpected discovery was made when voice acoustics alone approached the accuracy of the best multimodal model. This finding highlights the potential of voice in effectively revealing affective states, particularly those associated with OCD and comorbid depression. The inherent dynamism of the voice, coupled with its connection to the vagus nerve—the longest nerve in the autonomic nervous system and the primary nerve of the parasympathetic nervous system—renders it well-suited to capturing variations in arousal and stress levels. Notably, voice has been recognized as a reliable predictor of comorbid depression. For face, head, and gaze, it is worth noting that dynamic measures (e.g., velocity) contributed more than static ones (e.g., action unit duration). Multimodal models afforded increased prediction with the potential advantage of greater robustness to signal loss.

OCD and comorbid depression share some but not all features. For OCD five acoustic fearures were among the *top-k* features. For comorbid depression four of these were among the *top-k* as well. Four additional acoustic features were among the *top-k* for comorbid depression but not for OCD. For OCD none of the head and face dynamics features were among the *top-k* but for comorbid depression three head and face dynamics features were among the *top-k*. For OCD, one action unit feature was among the *top-k* but none for comorbid depression. Given the high correlation between OCD and comorbid depression, we must be cautious in interpreting these results; they do, however, suggest both similarities and differences between OCD and comorbid depression.

In behavioral and clinical science, explanation has historically been more important than prediction. Especially when dealing with health, explanation is critical. Recent work has called for an integration of explanation and prediction [124]. Our models were informative in detecting relative contribution of each modality and of key features within each modality.

A study limitation was the small number of participants. The participant pool from which to draw was small. Participants had to meet stringent criteria for severe and chronic treat-resistant DBS, additional psychiatric and medical criteria, opt for surgical implantation of electrodes deep in their brain, and participate in an 18-month trial. The within/participants (longitudinal) design with up to 8 assessments from each participant provided some offset to the limited numbers of participants. Supporting the validity of the findings was the consistency of the findings for depression. Comparable to previous research that had access to larger samples of participants, our findings for comorbid depression were quite consistent. For OCD and total energy delivered to the DBS, comparative data are unavailable. OCD and DBS energy are new research topics in affective computing.

To compare alternative algorithms or approaches by different investigators, access to common databases is essential. This is a problem when data are personally identifiable and involve sensitive topics, such as psychopathology. Audio and video as well as text may include personally identifiable information that is difficult to eliminate without impairing value of the data. In part for this reason, DAIC-WOZ has been one of the few depression databases made available to qualified researchers [7]. We will seek IRB permission to distribute de-identified and anonymized feature data from this clinical study for use by other researchers.

5 Conclusion

An unobtrusive, multimodal regressor based on open-ended interviews measured severity of OCD, severity of comorbid depression, and TEED over the course of a clinical trial for treatment resistant OCD. The regressor achieved strong consistency with state-of-the art clinical measures. With further validation, the proposed system could greatly reduce subjective variation in clinical judgment within- and between clinicians and eliminate drift over time in assessments for refractory OCD. An unexpected finding was the strength of acoustic features in inferring symptom severity and TEED. Facial action units and head and face dynamics contributed further predictive power. Linguistic features contributed relatively little. A key contributor to the modeling results was use of SHAP reduction in selecting most informative features. For the longitudinal data considered here, feature selection consistently improved performance. Inclusion of a mixed-effect for participants further contributed to improved performance for OCD and comorbid depression but not for TEED.

Acknowledgments

This research was supported in part by U.S. National Institutes of Health awards UH3 NS100549 and MH096951 and by the McNair Foundation. DBS devices were donated by Medtronic as part of the BRAIN Initiative Public-Private Partnership Program.

REFERENCES

[1] American Psychiatric Association, *DSM-5*, Washington, DC, 2015.

[2] S. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, and G. Parker, "Depression detection model interpretation via feature selection methods," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 133–151, 2023.

[3] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, no. C, p. 10–49, jul 2015. [Online]. Available: https://doi.org/10.1016/j.specom.2015.03.004

[4] H. Dibeklioglu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, 2018.

[5] M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu, "A multimodal fusion model with multi-level attention mechanism for depression detection," *Biomedical Signal Processing and Control*, vol. 82, p. 104561, 2023.

[6] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1–8, 2013.

[7] U. Arioz, U. Smrke, N. Plohl, and I. Mlakar, "Scoping review on the multimodal classification of depression and experimental study on existing multimodal models," *Diagnostics*, vol. 12, no. 11, 2022. [Online]. Available: https://www.mdpi.com/2075-4418/12/11/2683

[8] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches," *Sensors*, vol. 24, no. 2, p. 348, 2024.

[9] J. E.-C. Marjolein Fokkema and M. Wolpert, "Generalized linear mixed-model (glmm) trees: A flexible decision-tree method for multilevel and longitudinal data," *Psychotherapy Research*, vol. 31, no. 3, pp. 329–341, 2021, pMID: 32602811. [Online]. Available: https://doi.org/10.1080/10503307.2020.1785037

[10] R. A. Lewis, A. Ghandeharioun, S. Fedor, P. Pedrelli, R. Picard, and D. Mischoulon, "Mixed effects random forests for personalised predictions of clinical depression severity," *arXiv preprint arXiv:2301.09815*, 2023.

[11] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 3–10. [Online]. Available: https://doi.org/10.1145/2512530.2512533

[12] E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 13, no. 2, pp. 238–241, 1951.

[13] D. American Psychiatric Association, A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association Washington, DC, 2013, vol. 5, no. 5.

[14] L. C. Quarantini, A. R. Torres, A. S. Sampaio, V. Fossaluza, M. A. de Mathis, M. C. Do Rosário, L. F. Fontenelle, Y. A. Ferrão, A. V. Cordioli, K. Petribu *et al.*, "Comorbid major depression in obsessive-compulsive disorder patients," *Comprehensive psychiatry*, vol. 52, no. 4, pp. 386–393, 2011.

[15] S. A. Sheth and H. S. Mayberg, "Deep Brain Stimulation for Obsessive-Compulsive Disorder and Depression," *Annual Review of Neuroscience*, vol. 46, pp. 341–358, Jul. 2023.

[16] R. D. Crino and G. Andrews, "Obsessive-Compulsive Disorder and Axis I Comorbidity," *Journal of Anxiety Disorders*, 1996.

[17] T. Overbeek, K. Schruers, and E. Griez, "Comorbidity of Obsessive-Compulsive Disorder and Depression: Prevalence, Symptom Severity, and Treatment Effect," *The Journal of Clinical Psychiatry*, 2002.

[18] L. Bellodi, G. Sciuto, G. Diaferia, P. Ronchi, and E. Smeraldi, "Psychiatric disorders in the families of patients with obsessive-compulsive disorder," *Psychiatry Research*, 1992.

[19] U. Demal, G. Lenz, A. Mayrhofer, H. G. Zapotoczky, and W. Zitterl, "Obsessive-compulsive disorder and depression. A retrospective study on course and interaction," *Psychopathology*, 1993.

[20] R. J. Romanelli, F. M. Wu, R. Gamba, R. Mojtabai, and J. B. Segal, "Behavioral therapy and serotonin reu Behavioral therapy and serotonin reuptake inhibitor pharmacotherapy in the treatment of obsessive-compulsive disorder: a systematic review and meta-analysis of head-to-head randomized controlled trials," *Depression and Anxiety*, vol. 31, no. 8, pp. 641–652, 8 2014.

[21] L.-G. Öst, A. Havnen, B. Hansen, and G. Kvale, "Cognitive behavioral treatments of obsessive–compulsive disorder. A systematic review and meta-analysis of studies published 1993–2014," *Clinical Psychology Review*, vol. 40, pp. 156–169, 8 2015.

[22] Y. C. Janardhan Reddy, A. S. Sundar, J. C. Narayanaswamy, and S. B. Math, "Clinical practice guidelines for obsessive-compulsive disorder," *Indian J Psychiatry*, vol. 59, no. Suppl 1, pp. S74–s90, 2017.

[23] P. J. Karas, S. Lee, J. Jimenez-Shahed, W. K. Goodman, A. Viswanathan, and S. A. Sheth, "Deep brain stimulation for obsessive compulsive disorder: evolution of surgical stimulation target parallels changing model of dysfunctional brain circuits," *Frontiers in neuroscience*, p. 998, 2019.

[24] A. Rădulescu, J. Herron, C. Kennedy, and A. Scimemi, "Global and local excitation and inhibition shape the dynamics of the cortico-striatal-thalamo-cortical pathway," *Scientific Reports*, vol. 7, no. 1, p. 7608, 2017.

[25] R. Gadot, R. Najera, S. Hirani, A. Anand, E. Storch, W. K. Goodman, B. Shofty, and S. A. Sheth, "Efficacy of deep brain stimulation for treatment-resistant obsessive-compulsive disorder: systematic review and meta-analysis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 93, no. 11, pp. 1166–1173, 2022.

[26] B. Beebe and L. J. Gerstman, "The "packaging" of maternal stimulation in relation to infant facial-visual engagement: A case study at four months," *Merrill-Palmer Quarterly of Behavior and Development*, vol. 26, no. 4, pp. 321–339, 1980.

[27] M. V. McCall, P. Riva-Possea, S. J. Garlow, H. S. Mayberg, and A. L. Crowell, "Analyzing non-verbal behavior throughout recovery in a sample of depressed patients receiving deep brain stimulation," *Neurology, Psychiatry, and Brain Research*, vol. 37, pp. 33–40, 2020.

[28] R. O. Cotes, M. Boazak, E. Griner, Z. Jiang, B. Kim, W. Bremer, S. Seyedi, A. B. Rad, and G. D. Clifford, "Multimodal assessment of schizophrenia and depression utilizing video, acoustic, locomotor, electroencephalographic, and heart rate technology: protocol for an observational study," *JMIR Research Protocols*, vol. 11 —, no. 7, p. e36417, 2022.

[29] M. Bilalpur, S. Hinduja, L. A. Cariola, L. B. Sheeber, N. Alien, L. A. Jeni, L.-P. Morency, and J. F. Cohn, "Multimodal feature selection for detecting mothers' depression in dyadic interactions with their adolescent offspring," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023, pp. 1–8.

[30] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system: Research Nexus," in *Network Research Information, Salt Lake City, UT.*, 2002.

[31] A. S. Cowen and D. Keltner, "Universal facial expressions uncovered in art of the ancient americas: A computational approach," *Science advances*, vol. 6, no. 34, p. eabb1005, 2020.

[32] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[33] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional expression: Advances in basic emotion theory," *Journal of nonverbal behavior*, vol. 43, pp. 133–160, 2019.

[34] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures." *Emotion*, vol. 18, no. 1, p. 75, 2018.

[35] R. E. Mattson, R. D. Rogge, M. D. Johnson, E. K. Davidson, and F. D. Fincham, "The positive and negative semantic dimensions of relationship satisfaction," *Personal Relationships*, vol. 20, no. 2, pp. 328–355, 2013.

[36] I. O. Ertugrul, L. A. Jeni, W. Ding, and J. F. Cohn, "AFAR: A deep learning based tool for automated facial affect recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 5 2019, pp. 1–1.

[37] I. O. Ertugrul, J. F. Cohn, L. A. Jeni, Z. Zhang, L. Yin, and Q. Ji, "Crossing domains for AU coding: perspectives, approaches, and measures," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

[38] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[39] N. I. Technology, "Facereader v6.1," Report, 2015.

[40] Z. Hammal, J. F. Cohn, and D. T. George, "Interpersonal coordination of headmotion in distressed couples," *IEEE transactions on affective computing*, vol. 5, no. 2, pp. 155–167, 2014.

[41] Z. Hammal, J. F. Cohn, C. Heike, and M. L. Speltz, "Automatic

measurement of head and facial movement for analysis and detection of infants' positive and negative affect," *Frontiers in ICT*, vol. 2, 12 2015.

[42] M. Gavrilescu and N. Vizireanu, "Predicting depression, anxiety, and stress levels from videos using the facial action coding system," *Sensors*, vol. 19, no. 17, p. 3693, 8 2019.

[43] T.-H. Yang, C.-H. Wu, M.-H. Su, and C.-C. Chang, "Detection of mood disorder using modulation spectrum of facial action unit profiles," in *2016 International Conference on Orange Technologies (ICOT)*. IEEE, 12 2016, pp. 5–8.

[44] K. B. Martin, Z. Hammal, G. Ren, J. F. Cohn, J. Cassell, M. Ogihara, J. C. Britton, A. Gutierrez, and D. S. Messinger, "Objective measurement of head movement differences in children with and without autism spectrum disorder," *Molecular Autism*, vol. 9, no. 1, p. 14, 12 2018.

[45] Y. Ding, I. Onal Ertugrul, A. Darzi, N. Provenza, L. A. Jeni, D. Borton, W. Goodman, and J. Cohn, "Automated detection of optimal DBS Ddevice settings," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*. New York, NY, USA: ACM, 10 2020, pp. 354–356.

[46] A. Darzi, N. R. Provenza, L. A. Jeni, D. A. Borton, S. A. Sheth, W. K. Goodman, and J. F. Cohn, "Facial action units and head dynamics in longitudinal interviews reveal ocd and depression severity and dbs energy," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–6.

[47] J. Sundberg, S. Patel, E. Bjorkner, and K. R. Scherer, "Interdependencies among voice source parameters in emotional speech," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 162–174, 2011.

[48] D. T. Cordaro, D. Keltner, S. Tshering, D. Wangchuk, and L. M. Flynn, "The voice conveys emotion in ten globalized cultures and one remote village in bhutan." *Emotion*, vol. 16, no. 1, p. 117, 2016.

[49] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 582–596, 2009.

[50] M. Alpert, E. R. Pouget, and R. R. Silva, "Reflections of depression in acoustic measures of the patient's speech," *Journal of affective disorders*, vol. 66, no. 1, pp. 59–69, 2001.

[51] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[52] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE transactions on affective computing*, vol. 4, no. 2, pp. 142–150, 2012.

[53] T. Özseven, M. Düğenci, A. Doruk, and H. I. Kahraman, "Voice traces of anxiety: acoustic parameters affected by anxiety disorder," *Archives of Acoustics*, pp. 625–636, 2018.

[54] S. Scherer, G. Stratou, and L.-P. Morency, "Audiovisual behavior descriptors for depression assessment," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 135–140.

[55] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[56] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 9411–9457, 2021.

[57] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in parkinson's disease using a speech-to-text system," *IEEE Access*, vol. 5, pp. 22 199–22 208, 2017.

[58] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[59] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[60] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[61] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[62] C. Manning and H. Schutze, *Foundations of statistical natural language processing*. MIT press, 1999.

[63] A. S. Cohen, K. R. Mitchell, and B. Elvevåg, "What do we really know about blunted vocal affect and alogia? a meta-analysis of objective assessments," *Schizophrenia research*, vol. 159, no. 2-3, pp. 533–538, 2014.

[64] P. Baki, H. Kaya, H. Güleç, A. A. Salah *et al.*, "A multimodal approach for mania level prediction in bipolar disorder," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2119–2131, 2022.

[65] N. J. Carson, B. Mullin, M. J. Sanchez, F. Lu, K. Yang, M. Menezes, and B. L. Cook, "Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records," *PloS one*, vol. 14, no. 2, p. e0211116, 2019.

[66] M.-H. Metzger, N. Tvardik, Q. Gicquel, C. Bouvry, E. Poulet, and V. Potinet-Pagliaroli, "Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a french pilot study," *International journal of methods in psychiatric research*, vol. 26, no. 2, p. e1522, 2017.

[67] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural language processing of social media as screening for suicide risk," *Biomedical informatics insights*, vol. 10, p. 1178222618792860, 2018.

[68] A. Bittar, S. Velupillai, A. Roberts, and R. Dutta, "Text classification to inform suicide risk assessment in electronic health records." in *MedInfo*, 2019, pp. 40–44.

[69] M. Tanana, K. A. Hallgren, Z. E. Imel, D. C. Atkins, and V. Srikumar, "A comparison of natural language processing methods for automated coding of motivational interviewing," *Journal of substance abuse treatment*, vol. 65, pp. 43–50, 2016.

[70] M. J. Baggott, M. G. Kirkpatrick, G. Bedi, and H. de Wit, "Intimate insight: Mdma changes how people talk about significant others," *Journal of Psychopharmacology*, vol. 29, no. 6, pp. 669–677, 2015.

[71] D. To, B. Sharma, N. Karnik, C. Joyce, D. Dligach, and M. Afshar, "Validation of an alcohol misuse classifier in hospitalized patients," *Alcohol*, vol. 84, pp. 49–55, 2020.

[72] M. Hoogendoorn, T. Berger, A. Schulz, T. Stolz, and P. Szolovits, "Predicting social anxiety treatment outcome based on therapeutic email conversations," *IEEE journal of biomedical and health informatics*, vol. 21, no. 5, pp. 1449–1459, 2016.

[73] R. Patel, T. Lloyd, R. Jackson, M. Ball, H. Shetty, M. Broadbent, J. R. Geddes, R. Stewart, P. McGuire, and M. Taylor, "Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes," *BMJ open*, vol. 5, no. 5, p. e007504, 2015.

[74] T. Banerjee, M. Kollada, P. Gersberg, O. Rodriguez, J. Tiller, A. E. Jaffe, and J. Reynders, "Predicting mood disorder symptoms with remotely collected videos using an interpretable multimodal dynamic attention fusion network," *arXiv preprint arXiv:2109.03029*, 2021.

[75] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: The effect of gender," *Journal on Multimodal User Interfaces*, 2014.

[76] pp. 11–18, 2014.

[77] Z. Huang, J. Epps, D. Joachim, and M. Chen, "Depression detection from short utterances via diverse smartphones in natural environmental conditions," *Interspeech*, 2018.

[78] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. d. Santos, "Detection of major depressive disorder, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning," *Research on Biomedical Engineering*, vol. 38, p. 813–829, 2022.

[79] K. Schultebraucks, V. Yadav, A. Y. Shalev, G. A. Bonanno, and I. R. Galatzer-Levy, "Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood," *Psychological Medicine*, vol. 52, no. 5, p. 957–967, 2022.

[80] C. R. Marmar, A. D. Brown, M. Qian, E. Laska, C. Siegel, M. Li, D. Abu-Amara, A. Tsiartas, C. Richey, J. Smith, B. Knoth, and D. Vergyri, "Speech-based markers for posttraumatic stress disorder in US veterans," *Depression and Anxiety*, vol. 36, no. 7, pp. 607–616, Jul. 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/da.22890

[81] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal on Multimodal User Interfaces*, vol. 7, pp. 217–228, 2013.

[82] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 89–96.

[83] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, "Audio based depression detection using convolutional autoencoder," *Expert Systems with Applications*, vol. 189, p. 116076, 2022.

[84] Z. Zhang, W. Lin, M. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 344–350.

[85] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 319–330, 2001.

[86] Y. Zhou, X. Yao, W. Han, Y. Wang, Z. Li, and Y. Li, "Distinguishing apathy and depression in older adults with mild cognitive impairment using text, audio, and video based on multiclass classification and shapely additive explanations," *International Journal of Geriatric Psychiatry*, vol. 37, no. 11, 2022.

[87] J. J. Hox, *Multilevel analysis: Techniques and applications*. New York, NY: Routledge, 2010.

[88] B. G. Tabachnick and L. S. Fidell, *Multilevel liniear modeling*, 5th ed., 2007, book section 15, pp. 781–857.

[89] E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society, Series B*, vol. 13, p. 238–241, 1951.

[90] A. T. Beck, R. A. Steer, and G. Brown, *Manual for the Beck Depression Inventory-II*. San Antonio: Psychological Corporation, 1996.

[91] E. A. Storch, S. A. Rasmussen, L. H. Price, M. J. Larson, T. K. Murphy, and W. K. Goodman, "Development and psychometric evaluation of the Yale–Brown Obsessive-Compulsive Scale—Second Edition." *Psychological Assessment*, vol. 22, no. 2, pp. 223–232, 2010.

[92] M. D. McAuley, "Incorrect calculation of total electrical energy delivered by a deep brain stimulator," *Brain Stimulation*, vol. 13, no. 5, pp. 1414–1415, 9 2020.

[93] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D video for real-time use," *Image and Vision Computing*, 2017.

[94] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[95] J. M. Girard, W.-S. Chu, L. A. Jeni, and J. F. Cohn, "Sayette group formation task (gft) spontaneous facial expression database," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 581–588.

[96] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package)," *Neurocomputing*, vol. 307, pp. 72–77, 2018.

[97] C. Dewi, R.-C. Chen, X. Jiang, and H. Yu, "Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks," *PeerJ Computer Science*, vol. 8, p. e943, 2022.

[98] "TranscribeMe! - fast & accurate human transcription services." [Online]. Available: https://www.transcribeme.com/

[99] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: trainable text-speech alignment Using kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[100] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[101] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[102] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[103] M. Tasnim and J. Novikova, "Cost-effective models for detecting depression from speech," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2022, pp. 1687–1694.

[104] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/lio2.354

[105] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller, "Enhancing Speech-Based Depression Detection Through Gender Dependent Vowel-Level Formant Features," in *Artificial Intelligence in Medicine*, ser. Lecture Notes in Computer Science, A. ten Teije, C. Popow, J. H. Holmes, and L. Sacchi, Eds. Cham: Springer International Publishing, 2017, pp. 209–214.

[106] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 524–543, 2021.

[107] F. Haider, S. Pollak, P. Albert, and S. Luz, "Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods," *Computer Speech Language*, vol. 65, p. 101119, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300528

[108] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[109] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[110] H. Wu, *Nonparametric regression methods for longitudinal data analysis [mixed-effects modeling approaches]*, ser. Wiley series in probability and statistics. Hoboken, N.J: Wiley-Interscience, 2006.

[111] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed effects regression trees for clustered data," *Statistics & Probability Letters*, vol. 81, no. 4, pp. 451–459, Apr. 2011.

[112] L. S. Shapley, "A value for n-person games," *Classics in game theory*, vol. 69, 1997.

[113] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[114] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[115] I. Kim, S. Lee, Y. Kim, H. Namkoong, and S. Kim, "A probabilistic model for pathway-guided gene set selection," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 2733–2740.

[116] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, p. 307, Dec. 2008. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-307

[117] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, p. 80, Dec. 1945. [Online]. Available: https://www.jstor.org/stable/10.2307/3001968?origin=crossref

[118] S. Alghowinem, T. Gedeon, R. Goecke, J. F. Cohn, and G. Parker, "Interpretation of depression detection models via feature selection methods," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 133–152, 2023.

[119] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 73–80. [Online]. Available: https://doi.org/10.1145/2808196.2811641

[120] R. Rosenthal, "Conducting Judgment Studies:," in *The New Handbook of Methods in Nonverbal Behavior Research*, J. Harrigan, R. Rosenthal, and K. Scherer, Eds. Oxford University Press, Mar. 2008, pp. 199–234. [Online]. Available: https://academic.oup.com/book/25991/chapter/193835959

[121] A. J. D. Macdonald and A. J. B. Fugard, "Routine mental health outcome measurement in the UK," *International Review of Psychiatry (Abingdon, England)*, vol. 27, no. 4, pp. 306–319, 2015.

[122] "Brain Behavior Quantification & Synchronization Workshop," https://event.roseliassociates.com/bbqs-workshop, 2023, Accessed on 21st May 2023.

[123] N. R. Provenza, E. R. Matteson, A. B. Allawala, A. Barrios-Anderson, S. A. Sheth, A. Viswanathan, E. McIngvale, E. A. Storch, M. J. Frank, N. C. R. McLaughlin, J. F. Cohn, W. K. Goodman, and D. A. Borton, "The case for adaptive neuromodulation to treat severe intractable mental disorders," *Frontiers in Neuroscience*, vol. 13, no. 152, 2 2019.

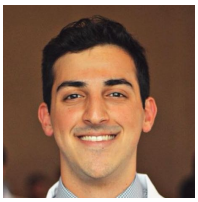
[124] J. M. Hofman, D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, A. Vespignani, and T. Yarkoni, "Integrating explanation and prediction in computational social science," *Nature*, vol. 595, pp. 181–188, 2021.

Saurabh Hinduja is a Post Doctorate Research Associate at the Affect Analysis Group, University of Pittsburgh, PA, USA. He received his PhD in Computer Science from University of South Florida. He has an MBA from the Symbiosis Center for Management and Human Resource Development (India) and a bachelors degree in engineering from the Birla Center for Management and Human Resource Development (India). His areas of interests include affective computing, artificial intelligence and machine learning. His work is in understanding contextual and self reported emotions. He is a member of the IEEE.

Ali Darzi is a visiting scholar and a former Postdoctorate Research Associate at the Affect Analysis Group, University of Pittsburgh, PA, USA. He received his Ph.D. in Electrical Engineering from the University of Wyoming, where his research focused on psychophysiology, rehabilitation robotics, and human-machine interaction. His interests lie in the fields of multimodal affective computing, digital signal and image processing, statistics, and machine learning. His current work revolves around the application of affective computing to develop a closed-loop adaptation paradigm for machines.

Itir Onal Ertugrul is an Assistant Professor at Social and Affective Computing Group, Department of Information and Computing Sciences at Utrecht University. Prior to joining Utrecht University, she was an Assistant Professor at Tilburg University, a postdoctoral researcher at the Robotics Institute at Carnegie Mellon University and Affect Analysis Group at University of Pittsburgh. She received B.Sc., M.Sc. and Ph.D. degrees from the Department of Computer Engineering at Middle East Technical University. Her research interests are in the broad areas of computer vision, machine learning, and affective computing, with a specific focus on automated analysis and synthesis of facial actions to understand human behavior, emotion, pain, and psychopathology.

Nicole Provenza is an assistant professor of Neurosurgery at Baylor College of Medicine (BCM). She received her Ph.D. in biomedical engineering from Brown University and completed her postdoctoral fellowship at BCM. Provenza's research focuses on the neurophysiology underlying cognition and emotion and the effects of neuromodulation on neural activity and behavior. Her work integrates neural activity and deep phenotyping approaches to inform neural signatures underlying real-world functional deficits in cognitive and emotional disorders.

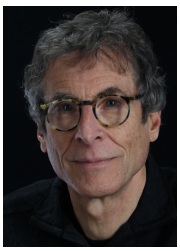


Ron Gadot is a medical student at Baylor College of Medicine and also a teaching assistant for the nervous system course at Baylor College of Medicine1. He is also affiliated with the Department of Neurosurgery at Baylor College of Medicine.

Eric A. Storch is a professor and the McIngvale Presidential Endowed Chair in the Department of Psychiatry and Behavioral Sciences at Baylor College of Medicine. He specializes in the cognitive-behavioral treatment of adult and childhood OCD, as well as other anxiety and OC-related disorders. His research interests involve the presentation and mechanisms and treatment of disorders, with a particular focus on driving innovation in the treatment of mental health disorders through the integration of academic research, evidence-based practice, and novel technological approaches.

Sameer A. Sheth is an Associate Professor of Neurosurgery and Vice-Chair of Clinical Research at Baylor College of Medicine. He specializes in the treatment of patients with Movement Disorders, Epilepsy, Brain Tumors, Trigeminal Neuralgia, Hydrocephalus, and certain Psychiatric Disorders. His translational research interests include developing and studying neuromodulation techniques for emerging neuropsychiatric conditions, including OCD, depression, addiction, schizophrenia.

Wayne K. Goodman is the D.C and Irene Ellwood Professor and chair of the Menninger Department of Psychiatry and Behavioral Sciences at Baylor College of Medicine. He is the principal developer, along with his colleagues, of the Yale-Brown Obsessive Compulsive Scale (Y-BOCS), which is considered to be the gold standard for assessing OCD. His research interests include OCD, deep brain stimulation, depression, and habenula.



Jeffrey F. Cohn is a professor of psychology, psychiatry, and intelligent systems at the University of Pittsburgh. He has led interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis of facial expression, body motion, and prosody and applied those tools to research in human emotion, interpersonal processes, social development, and psychopathology. He has co-developed influential databases (Cohn-Kanade, MultiPIE DISFA, Pain Archive, and the BP4D series), co-edited special issues on facial expression analysis, and chaired international conferences in automatic face and gesture recognition, multimodal interaction, and affective computing.