Emotion Recognition in the Real World: Passively Collecting and Estimating Emotions from Natural Speech Data of Individuals with Bipolar Disorder

Emily Mower Provost, *Senior Member, IEEE,* Sarah H Sperry, James Tavernor, Steve Anderau, Anastasia Yocum, Melvin G McInnis

***Abstract*—Emotions provide critical information regarding a person's health and well-being. Therefore, the ability to track emotion and patterns in emotion over time could provide new opportunities in measuring health longitudinally. This is of particular importance for individuals with bipolar disorder (BD), where emotion dysregulation is a hallmark symptom of increasing mood severity. However, measuring emotions typically requires self-assessment, a willful action outside of one's daily routine. In this paper, we describe a novel approach for collecting real-world natural speech data from daily life and measuring emotions from these data. The approach combines a novel data collection pipeline and validated robust emotion recognition models. We describe a deployment of this pipeline that included parallel clinical and self-report measures of mood and self-reported measures of emotion. Finally, we present approaches to estimate clinical and self-reported mood measures using a combination of passive and self-reported emotion measures. The results demonstrate that both passive and self-reported measures of emotion contribute to our ability to accurately estimate mood symptom severity for individuals with BD.**

***Index Terms*—Modeling human emotion, Mood or core affect, Diagnosis or assessment, Bipolar disorder**

I. Introduction

EMOTIONS are core human elements that express internal states and convey vital information about personal health and well-being. Emotions, measured using survey-based methodology (ecological momentary assessment, EMA), can identify features indicative of illness- or sickness-related states and behaviors across a spectrum of mental illnesses [1], including anxiety [2], bipolar disorder (BD) [3], depression [4], anorexia [5], and schizophrenia [1]. Yet, EMA requires individuals to respond to surveys, which is difficult to sustain over time [6]. Thus, there is increasing enthusiasm for identifying methods that can measure emotion passively, permitting the capture of natural behaviors while avoiding survey burden. However, there are major gaps in our understanding of how to create passive emotion recognition pipelines that are suitable for real-world deployment. As a result, there are few existing passive speech-centered emotion recognition tools as well as strategies to validate emotion estimates from passively collected data. In this paper, we present a pilot deployment of a real-world emotion-centered data collection for individuals with bipolar disorder (BD) using automatic speech emotion recognition tools. The presented work provides a template for the creation of passive emotion-centered pipelines and methods to validate the resulting automatic estimates of emotion.

Speech patterns convey critical information about mental health [7]–[16]. Changes in speech patterns are associated with clinical changes in mental status (e.g., [8]), particularly for individuals with BD (e.g., [17]–[24]), including changes due to emotion variability [25]–[27]. The ability to measure changes in emotion could lead to new avenues in the measurement of mental health symptom severity. However, automation requires the existence of speech-centered automatic emotion recognition pipelines that are robust to real-world conditions.

Once a robust pipeline is developed, a second problem emerges: the validation of information derived from that pipeline. There is a natural inclination to rely on a dual deployment of EMA and passive measures. Yet, these approaches are often not aligned. EMA is self-report, while passive measures are developed using datasets labeled in a perception-of-other paradigm, in which an outside group of annotators rate their perception of another individual's expression of emotion. Previous work has found that perception-of-other labels perform poorly in self-report contexts [28]–[34]. Therefore, since EMA is self-report and passive measures are trained based on perception-of-other labels, EMA alone provides an inaccurate and overly pessimistic view of the performance of passive techniques. This points to a need for new approaches for the validation of passive emotion estimates, in conjunction with the development of these new robust measures.

We describe a new pipeline for real-world emotion recognition, known as *PRIORI* (predicting individual outcomes for rapid intervention) that enables real-world emotion measurement from speech data without requiring action outside of day-to-day life. The PRIORI smartphone app collects audio data unobtrusively from an individual's ambient environment. It processes data in the cloud and estimates emotion (valence/activation). It does not retain the recorded audio; the audio data are deleted immediately following processing. The deployment also includes EMA self-reported emotion and

E. Mower Provost is in the Department of Computer Science and Engineering, University of Michigan e-mail: emilykmp@umich.edu

S. Sperry is in the Department of Psychiatry, University of Michigan e-mail: sperrys@med.umich.edu

J. Tavernor is in the Department of Computer Science and Engineering, University of Michigan e-mail: tavernor@umich.edu

S. Anderau is in the Department of Psychiatry, University of Michigan e-mail: standera@med.umich.edu

A. Yocum is in the Department of Psychiatry, University of Michigan e-mail: sperrys@med.umich.edu

M. McInnis is in the Department of Psychiatry, University of Michigan e-mail: mmcinnis@med.umich.edu

mood data, in addition to validated clinical scales. These measures of emotion, both passive and EMA, are used to estimate the level of mood symptom severity. The first step is extracting emotion features from both passive and EMA measures. These features are at the day-level and include: mean, inertia, variability, and the presence of emotional anomalies. The day-level features are aggregated into weekly features and are then used to estimate the mood symptom severity using linear mixed effect models (LMEM).

The paper presents a novel Institutional Review Board (IRB)-approved pipeline to both record and analyze emotion data, longitudinally, from daily life. The results are based on data collected between October 20, 2022 and February 19, 2024 over 20 individuals. The findings suggest that features derived from passive and EMA emotion measures capture clinical and self-reported mood severity and that the model coefficients align with the presentation of depression and mania symptomatology. Future work will focus on the integration of real-world self-report classifiers capable of operating in unstructured domains.

II. Related Work

A. Emotion and Bipolar Disorder

BD is a lifetime psychiatric condition characterized by pathological mood swings that range from mania to depression. The clinical patterns include changes from healthy (euthymic) mood states to mania (high energy), depression (low energy), and mixed states (symptoms of both depression and mania), often with severe consequences at a personal, social, vocational, and medical level [35], [36]. The hallmark of BD is emotional dysregulation that leads to intense biphasic extremes of energy, activity, and mood. The inherent and chronic variability of emotion and mood among people with BD [37], [38] makes this illness ideal for study.

Emotion states are periods of coordinated changes in neurophysiological activation, motor expression, subjective feeling, and action tendencies in response to internal or external perturbation [39]. Emotion is considered in the context of the dimensional circumplex model of core affect [40], which decomposes emotion states according to quantitative ratings of valence (pleasant to unpleasant) and activation (calm to activated, sometimes referred to as arousal), rather than discrete emotion states (e.g., anger, joy, shame). Distinct patterns of aroused/energized emotion states have been implicated in the pathophysiology of BD [41], [42]. This is reflected in the recent change to the Diagnostic and Statistical Manual for Mental Disorders – 5th Edition (DSM-5) [43], which includes a requirement that individuals with hypomania and mania experience increases in energy and activity regardless of whether they feel euphoric or irritable.

B. Robust and Deployable Speech Emotion Recognition

Speech-centered measures of emotion are challenging due to the difficulty of working with these data in real-world scenarios as speech is modulated by unseen (unobserved) contextual factors in which the methods are developed (e.g., background noise sources, different reverberation properties, the presence of other people, differences in speaking styles). These factors modulate the acoustics of the data, with implications for emotion recognition systems [44], [45], speech recognition (e.g., [46]–[51]), and mood severity estimation tasks [20], [52]. Methods have been proposed to account for differences in speech due to recording conditions and microphone quality [20], [52], noise robustness [53]–[56], and approaches to adapt features and models for cross-corpus testing [45], [57]–[61].

C. Deployed Sensor Platforms

Major research efforts have focused on modeling the illness-related behavior of individuals with mental health conditions [62], including BD [18]–[20], [63]–[67], [67]–[70] and the behavior of individuals with schizophrenia [1], [66], [71]–[75]. The StudentLife study, a study of Dartmouth students to assess mental health from passive sensor data, provides an example for validating and deploying engineering technologies for real-world measurement [76]–[83].

III. Pipeline Description

The pipeline includes **data recording** to capture the ambient audio data and **cloud processing** to extract the emotional information (Figure 1). This pipeline and protocol have been approved by the IRB at the University of Michigan (U-M, HUM00197298) and the platform has been assessed by the U-M Information Assurance office, who have confirmed that the platform and current cloud-based emotion recognition approach are compliant with university security protocols.

A. App Behavior

The platform records audio for 30-seconds every 15 minutes throughout the day. We calculate the maximum amplitude of the recorded speech and consider speech likely to be present if the maximum amplitude exceeds 1000[1]. The data are asymmetrically encrypted. Only encrypted audio are stored on device and cannot be decrypted on device. If the signal's maximum amplitude exceeds the threshold, the encrypted data are securely transmitted to a secure server via either Wi-Fi or cellular networks. After the encrypted audio data are uploaded or if the threshold amplitude is not reached, the audio data are immediately deleted from the smartphone.

B. Cloud Processing

There are six steps in the cloud processing pipeline (Figure 1). We first outline the steps and then discuss certain steps in more detail. **Step 1**: The data are first segmented into regions of speech and silence using webrtcvad[2]. **Step 2**: The speech segments are transcribed using automatic speech recognition (ASR) from Microsoft Azure[3]. **Step 3**: The segments are

[1]This is the threshold implemented in the Android source code of the app. A conservative threshold was selected during app development to prioritize the collection of audio. Silence is later removed in the cloud processing step using voice activity detection.

[2]https://pypi.org/project/webrtcvad/, aggressiveness=0

[3]https://azure.microsoft.com/en-us/products/ai-services/speech-to-text

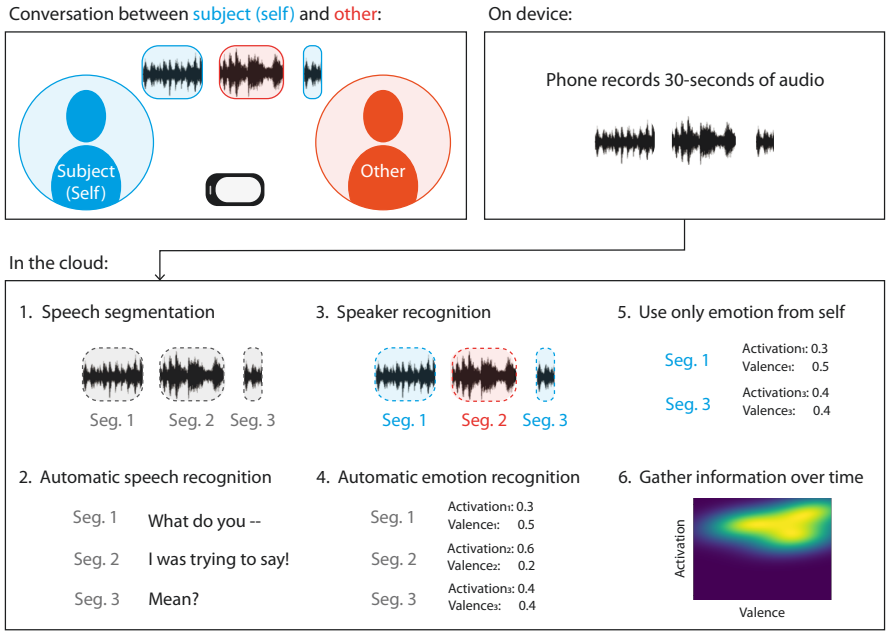

Fig. 1: The device records audio for 30-sec every 15-min and sends the audio to the cloud to determine the emotion (valence, activation). The audio data are asymmetrically encrypted and securely transmitted to a secure server. All data are then immediately deleted from the smartphone. In the cloud, the unencrypted data are deleted once processing concludes.

designated as belonging to the consented participant (vs. others in the participant's surroundings) using the Microsoft Azure speaker recognition tool[4]. We intentionally do not, and cannot, identify other speakers[5]. **Step 4**: Emotion recognition is then performed on the segments to extract measures of valence and activation (see Section IV). **Step 5**: The emotion inferences are associated with the consented participant or others in the participant's surroundings using the output of Step 3. **Step 6**: Features are extracted from the emotion inferences from the consented participant and the EMA surveys (Section V-B). These features are gathered over time (Section VI) and are then associated with mood (see Section VII).

C. Data Retained

Valence and activation estimates are stored with the probability that the speaker in the segment is the consented participant, the length of the segment, and the signal to noise ratio (SNR) of the segment (dB). The original audio, the extracted features, and estimated transcript are not saved.

IV. Emotion Recognition

A. Data

The emotion recognition algorithms are trained using MSP-Podcast, a dataset curated from publicly available podcast audio [84]. The dataset includes ratings of both valence and activation. We use the predefined training split and validation splits. We created transcripts using the Microsoft Azure ASR described in Section III-C.

[4]https://azure.microsoft.com/en-us/products/ai-services/speaker-recognition

[5]Note that in all experiments discussed in this paper, we use only inferences extracted from speech identified as being from the consented participant with probability greater than 0.5, the default setting for speaker recognition from Microsoft Azure.

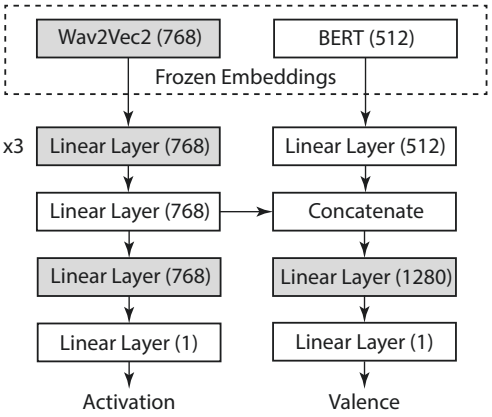

Fig. 2: The model architecture for the robust emotion approach. All linear layers have LayerNorm applied to their inputs, and all layers in gray have LeakyReLU and Dropout applied to their inputs.

B. Model

The emotion recognition model used in this study is based on our prior work [45]. The segmented speech and text data are encoded using large, pretrained transformers. We use Wav2Vec2 [85] acoustic features (they generally are more effective, compared to Mel-spectrogram features [56]) and BERT (Bidirectional Encoder Representations from Transformers) language features [86]. The BERT features are taken from the hidden state corresponding to the CLS (classification) token. The CLS token has LayerNorm and dropout (p=0.2) applied and the result is passed through a linear layer. The Wav2Vec2 features are calculated by taking the mean over the last dimension of the last hidden state.

The dual predictions include both valence and activation. Predicting valence and activation jointly, in a multitask context, provides regularization [87]. In addition to mean pooling

of Wav2Vec2, we apply leaky ReLU, LayerNorm, and dropout (p=0.2). We use four linear layers on top of the Wav2Vec2 features to allow for additional complexity and ensure the model can learn more general acoustic representations. Two linear layers are then used as prediction heads for both activation and valence. The first three layers after Wav2Vec2 features and the first layer in each prediction head apply LeakyReLU and dropout with probability 0.2. All linear layers apply LayerNorm to their inputs. Additionally, as lexical input does not improve activation performance when using Wav2Vec2 features, we do not apply the BERT input to the activation prediction [56]. Instead, lexical content is concatenated to the acoustic representation before the valence prediction head. This allows us to learn valence without degrading activation performance. The model is trained using early stopping with a patience of 15, batch size of 32, and optimized with stochastic gradient descent with a learning rate of 1e-3 using Lin's Concordance Correlation Coefficient (CCC) as the model's evaluation metric and loss function. The loss function and evaluation metric are those recommended in [56].

The models are made domain robust using the Gradient Episodic Memory (GEM) framework designed for continual learning [88]. The model tracks the most recently seen samples from a task and stores them in episodic memory. Instead of replaying the previously seen samples, GEM iterates over the memory samples calculating the loss and proposed weight updates during each training step. GEM compares the proposed weight update for each previous task to the current weight update. Considering these weights as vectors, GEM will prevent an update if the gradients conflict (the inner product is less than zero). When an update is presented, GEM solves a quadratic programming problem to find an update as close as possible to the original update that does not conflict with past task gradients. The advantage of GEM is that it avoids overfitting to memory samples while ensuring strong memory of the previous tasks. GEM has been used for automatic speech recognition to avoid retraining on complete datasets when new data is introduced to improve on total training time [89]. The system obtains 0.65-0.68 CCC for activation and 0.57-0.59 CCC for valence on IEMOCAP and MSP-Improv, two standard emotion recognition corpora, and exhibits accurate cross-corpus prediction [45].

V. Pilot Deployment

We have recruited twenty participants since May 2022, with 16 participants having sufficient data (68.75% BD I, 18.75% BD II, 12.50% Healthy Control (HC), Mean Age = 53.94, SD Age = 13.14, 62.50% female, 93.75% White). In the results that follow, we discuss data that have been collected from October 20, 2022 [6] to February 19, 2024.

A. Enrollment

Participants were included based on: 1) either a BD I or II diagnosis based on the Diagnostic and Statistical Manual IV (DSM-IV) criteria or 2) as healthy controls (HC, n=1), between the ages of 18 - 70, and use of an Android mobile phone. Exclusion criteria included: hazardous use of alcohol or drugs in the last three months, history of medical or neurological conditions known to chronically affect speech as determined by review of medical records and the clinical PI. All participants reviewed the IRB approved consent form with a study coordinator and provided their signature.

Participants' audio data are recorded during enrollment while they read a standardized text passage, the Caterpillar Passage [90]. These data are used as speaker's reference data for speaker recognition.

B. Protocol

Participants complete an EMA protocol one week per month while enrolled in the study. This involves a *measurement burst approach* that captures variability in self-reported emotion over the observation period while minimizing the overall burden to the participant.

EMA surveys were deployed via the *MyDataHelps* app (Care Evolution, Ann Arbor, MI), a HIPAA compliant and U-M approved mHealth application. EMA data are pushed each evening from the CareEvolution cloud-based server to our local servers through an SFTP protocol. At no point does any of the data transmitted between MyDataHelps and our server include identifying information except a coded identifier. During a burst week, emotion surveys are administered five times per day (9am, 12pm, 3pm, 6pm, 9pm). Consistent with the affective circumplex theory [91], participants are asked to self-rate their emotional valence on a Likert scale from -5 (unpleasant) to 5 (pleasant) and emotional activation on a Likert scale from -5 (sleepy, calm) to 5 (activated/jittery).

Self-ratings of mood are administered using the digital survey for mood in BD (*digiBP*), a brief 6-item survey that is validated in digital form to separately measure manic and depressive symptom severity [92]. The six items were selected from the Young Mania Rating Scale (YMRS) and the Hamilton Depression Rating Scale (HDRS) to be consistent with gold-standard clinician-rated measures. Three items measure depressive symptoms (depressed mood, fatigue, fidgeting), two measure manic symptoms (increased energy, rapid speech), and one item measures a symptom of both mania and depression (irritability). Each item is rated on an ordinal scale: 0=absent/normal, 1=mild, 2=moderate, 3=severe. Two scores, D and M, are computed to measure severity of depressive and manic symptoms, respectively. A factor analysis model confirmed that the survey items load onto two factors: "manic" (M) and "depressive" (D). Weekly averages of M and D scores were also able to explain significant variation in weekly scores from the YMRS (R^2 = 0.47) and HDRS (R^2 = 0.58). The digiBP survey has also been validated as a within-person measure for predicting a person's future M and D scores [92].

During the burst week, participants complete a self-report of mood twice daily (*digiAM* and *digiPM*), using the digiBP [92]–[94]. We take the average of these two self-reports to form a daily rating (*digiDay*). Outside of the burst week participants complete a weekly digiBP (*digiWeek*). We refer to

[6]This is when the activation classifier was changed to the one described in Section IV.

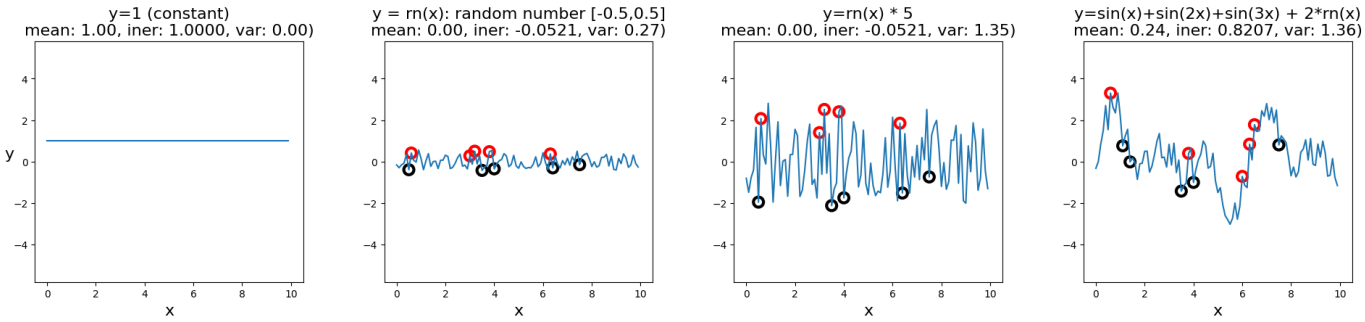

Fig. 3: Graphical depictions of mean, inertia ("iner", mean of the normalized auto-correlation over shift of size 1), and variability ("var", standard deviation) across four different signals. The left-most signal is constant, the inertia is 1. Next, we generate a random number, rn, for each value of x $y = rn(x)$. The inertia drops. Next, we multiply the random number by five: $y = 5 * rn(x)$. The inertia is unchanged by a scaling factor. Next, we add structure in the form of a sum of sine waves: $y = sin(x) + sin(2x) + sin(3x) + 2 * rn(x)$. The inertia increases due to the underlying structure. We show anomalous measurements that transition from a low value to a high value in red (large difference between subsequent measurements) and from a high value to a low value in black (small difference between subsequent measurements) circles.

Measure	Total	Mean	Standard Deviation
Total Days	2403	120.15	80.28
Unique Recordings	50364	2518.20	2281.35
Speech Segments	91787	4589.35	4319.08
EMA Ratings	5035	251.75	102.86
Daily digiBP	2227	111.35	42.67
Weekly digiBP	692	34.60	12.35
HDRS	125	6.25	1.92
YMRS	151	7.55	2.84

TABLE I: Total data collected. The mean and standard deviation are at the participant-level.

Survey	Passive	EMA	Passive + EMA
YMRS	88	124	77
HDRS	76	104	68
digiWeek_M	396	311	185
digiWeek_D	396	311	185
digiDay_M	607	988	605
digiDay_D	607	988	605

TABLE III: Dataset size for each target label considering availability of passive, EMA, or passive+EMA.

Survey	Within-Person	Between-Person
digiWeek_D	0.675	0.943
digiAM_D	0.643	0.897
digiPM_D	0.523	0.908
digiWeek_M	0.546	0.900
digiAM_M	0.514	0.894
digiPM_M	0.519	0.832

TABLE II: Omega reliability statistics for the digiBP metrics.

the mania subscale of the weekly digiBP as *digiWeek_M* and the depression subscale as *digiWeek_D*. The morning, evening, and daily measures follow the same naming convention (e.g., *digiDay_M*).

C. Weekly Mood-State Clinical Assessment

At the end of each EMA burst assessment week (see Section V-B), participants complete a telephone interview with a study clinician to assess symptom severity; manic symptoms via the YMRS and depressive symptoms via the HDRS.

D. Statistics of Recorded Data

We have collected 2,403 days of PRIORI recordings over the 20 participants. This includes 50,364 unique recordings and 91,787 speech segments. Please see Table I for additional statistics.

We first present omega reliability metrics (omega total) for the digiBP surveys, given the newness of this metric. These metrics were calculated using the reliabiliPy Python package [95]. The within- and between-person total omega for the weekly digiBP can be seen in Table II. This confirms that under the conditions of this study, the digiBP is considered highly reliable and a consistent method to assess mood states.

E. Creation of an Emotion-Mood Dataset

For each participant, data are included if they are recorded within seven days of any of the target mood measures (YMRS, HDRS, digiWeek_M, digiWeek_D, digiDay_M, digiDay_D). Days are processed only if both EMA and passive measures were collected, and there must be at least one such day within the seven days. This restriction permits direct comparability between EMA and passive measures, but does result in a smaller dataset compared to one requiring only EMA or only passive. In future work we will lift this restriction.

VI. Emotion Features

While a number of indices can be extracted from emotion time-series, it is the dynamic measures that have been shown to robustly predict BD symptoms [96]–[98]. These measures include: intensity of high activation negative and positive emotion (mean), standard deviation in reported emotion (variability), autocorrelation (inertia), and large acute shifts in valence and activation compared to one's own average (anomalies). Anomalies in high activation emotions, either negative or positive, predict the development of BD in those at risk [99], highlighting these dynamic indices as critical to the study of BD over-and-above mean levels. Inertia captures the

predictability of future measures, based on current measures, and variability captures the deviation from one's mean [97]. See Figure 3 for a visual depiction of the emotion features.

In all cases, the emotion features are calculated by aggregating all emotion measures (passive estimates, EMA) over a single day. Mean is calculated as the average valence or activation rating over the day, anomalies are identified by calculating the difference between successive measures, normalized by the median distance in time. We differentiate between upper anomalies (a positive difference, moving from a lower to a higher value) and lower anomalies (a negative difference, moving from a higher to a lower value). We consider measurements to be anomalous if they are above the 95th percentile (upper) or below the 5th percentile (lower), for that participant, based on practices in the field [97] (see Figure 3 for a depiction of lower, black, vs. upper, red, anomalies). Inertia is calculated by computing the mean of the normalized autocorrelation of the signal shifted by one. Intuitively, signals that are less random (more predictable based on past measurements), will have higher inertia (see Figure 3 for an illustrative example). Variability is calculated as the within-person standard deviation of the emotion measurements.

We refer to **passive features** as those extracted from the automatic estimates and **EMA features** as those extracted from the self-report EMA surveys. Each measure is calculated separately for passive and EMA estimates. The final feature vector for each of the survey targets (Table III) is the average over the seven-day window. There are 10 passive features and 10 EMA features (in each case, five for valence and five for activation).

VII. Analyses of Emotion and Clinical Mood Measures

The relationship between the emotion measures (EMA and passive) and clinical mood measures (YMRS, HDRS) is first studied by fitting a linear mixed effect model (LMEM), implemented in statsmodels [100], to the entire population. This model is not predictive, rather it demonstrates the ability of the features (daily mean, inertia, variability, anomaly) derived from the EMA and passive measurements to explain the variance observed in the mood measures. The emotion features are z-normalized prior to fitting each model and a random effect for gender is included[7]. There are relatively few clinical measures recorded for each participant. The requirement of collection of both EMA and passive measures on a given day and within seven days of a YMRS results in 77 weekly aggregates across all participants, and 68 for HDRS (vs. the relatively larger number of self-reported mood measures in the next section, 605).

In the first set of analyses, the model fit is quantified in terms of both marginal and conditional R^2 and Pearson's Correlation Coefficient (PCC). Marginal, R^2_m, is with respect to the fixed factors (e.g., passive mean valence) and conditional, R^2_c, is with respect to both the fixed and random factors (i.e., identity). For readability, we present only PCC results in the text. We refer the reader to Table IV for the full results.

LMEM models were fit first on the YMRS data. The first model uses only passive measures ($PCC = 0.37$). The second model uses only EMA data ($PCC = 0.49$). The model fit improves when modeling both types of data together ($PCC = 0.56$).

In the analysis with HDRS data, the first model again uses only passive measures ($PCC = 0.41$). The model is repeated using the EMA data ($PCC = 0.62$). The results again improve when modeling both types of data together ($PCC = 0.74$). For full results, including results by feature type (e.g., mean), see Table IV.

A. EMA, Passive, and Self-Report Mood Measures

The relationship between the emotion measures (EMA and passive) and self-reported daily mood measures (digiDay_M, digiDay_D) are first fitted using LMEM applied to the entire population, as discussed in the previous section. The emotion features are again z-normalized prior to fitting each model and a random effect for gender is included. See Table V for the standardized coefficients.

LMEM models were first fit for the digiDay_D data. The first LMEM is fit on only the passive features ($PCC = 0.29$). We find that both the combination of all passive features outperforms that of a single feature alone and that the EMA features have higher PCC, compared to the passive features ($PCC = 0.67$). The results show that the combination of passive and EMA features outperforms either approach alone across all metrics ($PCC = 0.70$). See Table IV for full results and Figure 4 for a visualization of the model fit on the passive and EMA data.

LMEM models were then fit for the digiDay_M data, first using passive, then EMA, then a combination of passive and EMA features. The LMEM fit with passive features has a $PCC = 0.34$. The combination of all passive features outperforms that of a single feature alone. The LMEM fit with EMA features has higher PCC, compared to the passive features ($PCC = 0.52$). Again, the combination of passive and EMA features outperforms either approach alone ($PCC = 0.59$). See Table IV for full results and Figure 5 for a visualization of the model fit on the passive and EMA data.

B. Participant-Specific Results

Thus far, the results presented have considered a dataset composed of all participants together. In the next analysis, the participant data are disaggregated into participant-specific sets. The LMEM is first evaluated in a data fitting capacity and then in a prediction capacity.

First, LMEM models were fit using all participants' given their passive and EMA measures. The LMEM is not retrained and is therefore not subject-independent at this point. The PCC between the LMEM output and the true digiDay_D is 0.60 $\pm$ 0.45 (mean $\pm$ standard deviation) and for digiDay_M is 0.63 $\pm$ 0.40 (Figures 6 and 8).

Next, LMEM models were run in a prediction capacity using a leave-one-participant-out approach. There are two sets of

[7]We use gender, rather than subject identity, to allow for a leave-one-subject-out analysis.

	Scale	Mean			Variability			Inertia			Upper Anomaly			Lower Anomaly			All		
		R_m^2	R_c^2	PCC	R_m^2	R_c^2	PCC	R_m^2	R_c^2	PCC	R_m^2	R_c^2	PCC	R_m^2	R_c^2	PCC	R_m^2	R_c^2	PCC
Passive	ymrs	0.07	0.25	0.18	0.06	0.11	0.26	0.10	0.17	0.33	0.04	0.07	0.22	0.02	0.07	0.16	0.16	0.35	0.37
	hdrs	0.02	0.02	0.14	0.04	0.04	0.19	0.02	0.02	0.16	0.03	0.03	0.18	0.03	0.03	0.17	0.09	0.54	0.41
	dD_D	0.00	0.29	0.06	0.03	0.04	0.17	0.01	0.29	0.12	0.00	0.03	0.00	0.00	0.03	0.04	0.05	0.35	0.29
	dD_M	0.00	0.06	0.08	0.01	0.05	0.15	0.01	0.07	0.11	0.00	0.07	0.07	0.06	0.12	0.24	0.07	0.34	0.34
EMA	ymrs	0.01	0.09	0.11	0.09	0.22	0.28	0.00	0.06	0.01	0.04	0.11	0.19	0.10	0.15	0.34	0.21	0.32	0.49
	hdrs	0.19	0.19	0.43	0.00	0.00	0.06	0.01	0.01	0.10	0.13	0.13	0.36	0.02	0.02	0.15	0.36	0.38	0.62
	dD_D	0.25	0.49	0.58	0.01	0.04	0.09	0.01	0.04	0.09	0.02	0.04	0.16	0.00	0.29	0.06	0.35	0.56	0.67
	dD_M	0.22	0.29	0.47	0.00	0.07	0.05	0.01	0.07	0.13	0.01	0.07	0.10	0.01	0.07	0.13	0.25	0.29	0.52
Combined	ymrs	0.09	0.30	0.20	0.12	0.22	0.35	0.10	0.17	0.33	0.07	0.11	0.29	0.10	0.15	0.34	0.28	0.45	0.56
	hdrs	0.21	0.26	0.45	0.04	0.04	0.21	0.04	0.04	0.20	0.14	0.14	0.39	0.07	0.07	0.27	0.47	0.54	0.74
	dD_D	0.25	0.49	0.58	0.03	0.05	0.20	0.02	0.05	0.14	0.02	0.04	0.16	0.00	0.29	0.08	0.40	0.58	0.70
	dD_M	0.22	0.28	0.48	0.01	0.06	0.16	0.02	0.08	0.17	0.01	0.07	0.13	0.06	0.12	0.26	0.26	0.48	0.59

TABLE IV: Marginal and conditional R^2 for LMEMs. Marginal R_m^2 is with respect to the fixed factors, while the conditional R_c^2 is with respect to both the fixed and random factors. The scale dD refers to digiDay (e.g., digiDay_D).

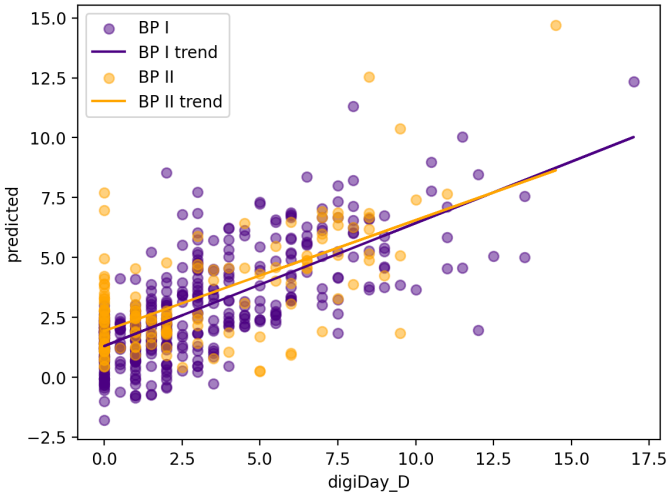

Fig. 4: The LMEM data fitting for digiDay_D. The x-axis represents the true value, the y-axis is the predicted value. The purple points are data from individuals with a BP I diagnosis, while the orange points are from individuals with a BP II diagnosis. The straight lines represent the trends for each group.

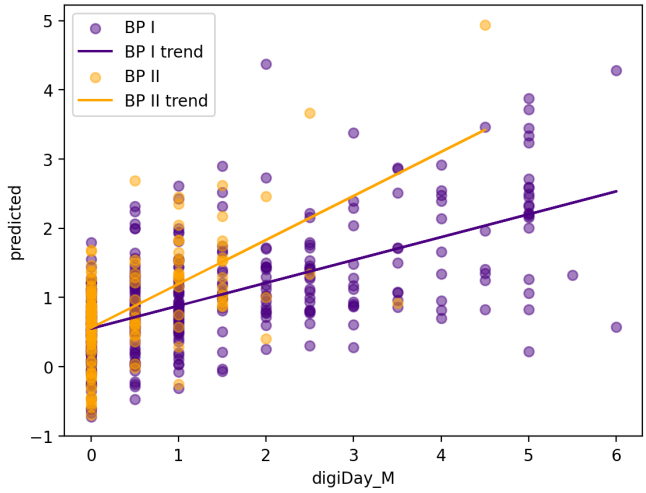

Fig. 5: The LMEM data fitting for digiDay_M. The x-axis represents the true value, the y-axis is the predicted value. The purple points are data from individuals with a BP I diagnosis, while the orange points are from individuals with a BP II diagnosis. The straight lines represent the trends for each group.

data: 1) the training data, containing $n-1$ participants and 2) the testing data, containing the final remaining participant. Each participant serves as a test participant. The LMEM is created using the training data and evaluated on the testing data. Therefore, there are n distinct LMEMs that are trained. The result is a participant-independent analysis (participants are not in both the training and the testing data). The PCC for digiDay_D is 0.56 ± 0.48 and for digiDay_M is 0.61 ± 0.42 (Figures 7 and 9).

C. Analysis of Linear Mixed Effect Models

Given the small sample size of this pilot study, analysis of the combined LMEM using all emotion features concentrated on 95% confidence intervals (CI) rather than p-values, 95% CI's that include 0 were considered non-significant. The range of the lower and upper bound of the 95% describes the confidence limits of the estimate. See Table V for full results.

1) Self-reported Mania: Greater variability in passive valence was associated with higher digiDay_M such that a 1-unit increase in variability was associated with a 0.22 standard deviation (SD) increase in digiDay_M score. Greater variability in passive activation was associated with a decrease in digiDay_M such that a 1-unit increase in variability of activation was associated with a 0.28 SD decrease in digiDay_M. A greater number of anomalies (lower: from high to low values) in passive activation were associated with higher mania scores. In terms of EMA, mean valence was negatively associated with mania scores such that a 1-unit increase in EMA valence was associated with a 0.64 SD decrease in mania. The opposite was true of EMA activation, a 1-unit increase in activation was associated with a 0.59 SD increase in mania scores.

2) Self-reported Depression: A 1-unit decrease in mean levels of passive valence was associated with a 0.37 SD increase in depression scores. A 1-unit increase in the variability of passive valence was associated with a 0.46 SD increase in depression scores. A 1-unit increase in the variability of passive activation was associated with a 0.58 SD decrease in depression scores. In addition, inertia of passive valence was associated with a 0.38 SD increase in depression scores.

feature	coef.	std. error	p-value	95% CI
digiDay_M				
passive val mean	*-0.10*	*0.06*	*0.064*	*-0.21 - 0.01*
passive act mean	0.07	0.06	0.252	-0.05 - 0.18
passive val vari	**0.22**	**0.08**	**0.003**	**0.08 - 0.37**
passive act vari	**-0.28**	**0.07**	**< 0.001**	**-0.42 - -0.13**
passive val iner	**0.18**	**0.05**	**< 0.001**	**0.08 - 0.29**
passive act iner	-0.04	0.06	0.475	-0.16 - 0.07
passive val anom upper	**-0.13**	**0.06**	**0.022**	**-0.24 - -0.02**
passive act anom upper	-0.01	0.06	0.820	-0.13 - 0.10
passive val anom lower	0.10	0.06	0.102	-0.02 - 0.21
passive act anom lower	**0.19**	**0.06**	**< 0.001**	**0.08 - 0.31**
ema val mean	**-0.64**	**0.06**	**< 0.001**	**-0.75 - -0.53**
ema act mean	**0.59**	**0.06**	**< 0.001**	**0.48 - 0.71**
ema val vari	0.06	0.08	0.440	-0.09 - 0.21
ema act vari	*0.12*	*0.07*	*0.084*	*-0.02 - 0.25*
ema val iner	**-0.15**	**0.06**	**0.013**	**-0.26 - -0.03**
ema act iner	*0.10*	*0.05*	*0.062*	*-0.01 - 0.21*
ema val anom upper	**-0.16**	**0.05**	**0.002**	**-0.26 - -0.06**
ema act anom upper	0.04	0.05	0.459	-0.06 - 0.14
ema val anom lower	**-0.11**	**0.06**	**0.050**	**-0.22 - -0.00**
ema act anom lower	0.04	0.05	0.444	-0.06 - 0.14
digiDay_D				
passive val mean	**-0.37**	**0.11**	**0.001**	**-0.58 - -0.15**
passive act mean	0.05	0.12	0.673	-0.18 - 0.28
passive val vari	**0.46**	**0.15**	**0.002**	**0.17 - 0.75**
passive act vari	**-0.58**	**0.15**	**< 0.001**	**-0.87 - -0.29**
passive val iner	**0.38**	**0.11**	**< 0.001**	**0.18 - 0.59**
passive act iner	0.15	0.12	0.212	-0.08 - 0.38
passive val anom upper	-0.04	0.11	0.744	-0.26 - 0.18
passive act anom upper	0.18	0.12	0.133	-0.05 - 0.41
passive val anom lower	**-0.32**	**0.12**	**0.006**	**-0.56 - -0.09**
passive act anom lower	**0.29**	**0.11**	**0.010**	**0.07 - 0.52**
ema val mean	**-2.25**	**0.11**	**< 0.001**	**-2.47 - -2.03**
ema act mean	**0.40**	**0.12**	**< 0.001**	**0.17 - 0.64**
ema val vari	-0.03	0.15	0.835	-0.32 - 0.26
ema act vari	**0.55**	**0.13**	**< 0.001**	**0.29 - 0.82**
ema val iner	**0.76**	**0.12**	**< 0.001**	**0.53 - 0.99**
ema act iner	0.18	0.11	0.101	-0.03 - 0.39
ema val anom upper	**-0.26**	**0.10**	**0.011**	**-0.47 - -0.06**
ema act anom upper	0.11	0.10	0.262	-0.08 - 0.31
ema val anom lower	**0.29**	**0.11**	**0.010**	**0.07 - 0.51**
ema act anom lower	0.02	0.10	0.860	-0.18 - 0.21

TABLE V: LMEM descriptions for self-reported mood. Bold: coefficients with p-value < 0.05, italics: coefficients with p-value < 0.1. CI: confidence interval.

A greater number of anomalies (from high to low scores) in valence was associated with a 0.32 SD decrease in depression whereas a greater number of anomalies in passive activation were associated with a 0.29 SD increase in depression. In terms of EMA, a 1-unit decrease in valence mean was associated with a 2.25 SD increase in depression scores. Higher EMA activation was associated with a 0.40 SD increase in depression. Similar to passive measures, greater inertia in valence was associated with 0.76 SD increase in depression. The opposite was observed for EMA anomalies: a greater number of high to low valence anomalies was associated with an increase in digiDay_D.

VIII. Relationship between EMA and Passive Measures

The relationship between EMA and passive estimates reinforce findings from the emotion recognition literature, speaking to the differences between self-report (EMA is self-report) and passive estimates [30], [31]. We observe a similar trend. The mean measures extracted from the passive and EMA data are not correlated across either valence (-0.00) or activation (-0.10). For the valence dimension, the variability feature is the most well-correlated across EMA/passive measures (0.13), compared to inertia (0.05), upper anomalies (-0.03), and lower anomalies (-0.13). There is a similar trend for the activation dimension (variability: 0.20, inertia: -0.00, upper anomalies: 0.15, lower anomalies: 0.07). If used as the sole validation, the results suggest that passive measures are not accurate. However, it is important to consider alternative validation approaches given the differences that exist between EMA (self-report) and passive measures (perception-of-other) and that insight into one's emotions can be affected differently based on current mood state [101].

IX. Discussion

The results suggest that EMA and passive measurements both contribute, but contribute differently, to the explanation of variance in mood symptom severity. We find that, measures derived from the passive features generally contributed more information to the modeling of clinical measures (YMRS, HDRS) compared to self-report measures (digiDay_M, digiDay_D).

In the modeling of symptom severity, measures derived from both the passive and EMA collections provide substantial insight into the variability of symptom severity. In the case of HDRS and digiDay_D, features derived from EMA are generally more effective for modeling depression symptom severity, compared to passive features. This is consistent with prior findings showing that emotional insight in depressive episodes is not impaired [102]. Yet, the combination of passive and EMA features leads to the highest performance. These findings are echoed in the prediction of mania symptom severity. Again, the EMA features provide a stronger signal, compared to those derived from the passive measures. However, again, we see that the combination of both passive and EMA features leads to the highest performance.

Overall, the preliminary findings support a growing body of literature that suggests that the temporal dynamics of emotional valence and activation is related to mood severity [96], [97] over and above mean levels. We observed that the variability of passive valence is associated with higher self-reported depression (digiDay_D) and mania (digiDay_M) scores. Conceptually, this suggests that higher mood scores are associated with greater variability in valence (e.g., greater reactivity or dysregulation) captured in speech. We further observered that increased variability of passive activation is associated with lower depression and mania scores. Conceptually, this suggests that lower variability in activation is associated with higher mood scores (e.g., consistently low activation in depression, consistently high activation in mania). However, we find that the variability of EMA activation is associated with higher depression, but not mania, scores. This may speak to a self-perception of high activation that may not be reflected in observable behavior. Additional research is needed to investigate whether this trend will be replicated across a larger sample. Fundamentally, these findings speak to the increasing attention paid to the role of activation and energy as a core feature of mood episodes in bipolar disorder [98], [103]–[105].

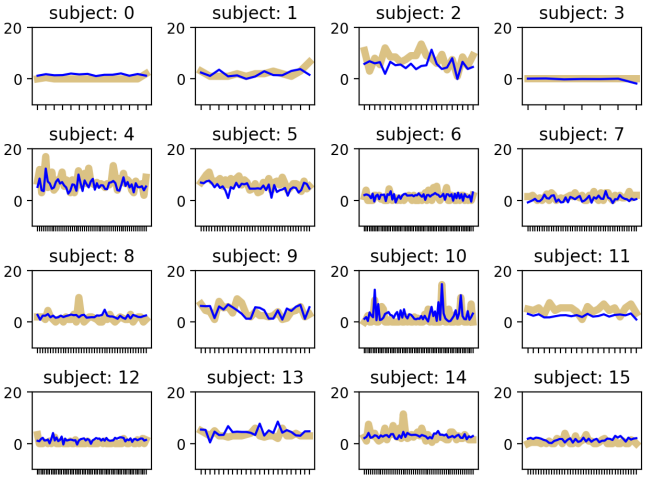

Fig. 6: The LMEM for digiDay_D, separated by subjects and ordered in time. The brown line is the self-reported ground truth. The blue line is the estimated value. These data were generated through **data fitting**.

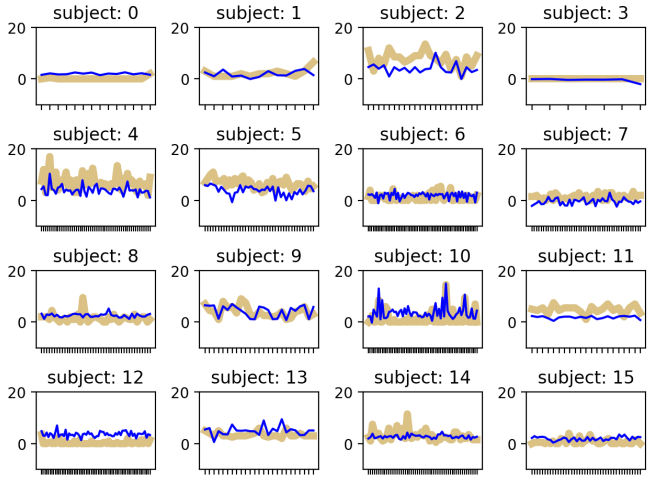

Fig. 7: The LMEM for digiDay_D, separated by subjects and ordered in time. The brown line is the self-reported ground truth. The blue line is the estimated value. These data were generated through **subject independent prediction**.

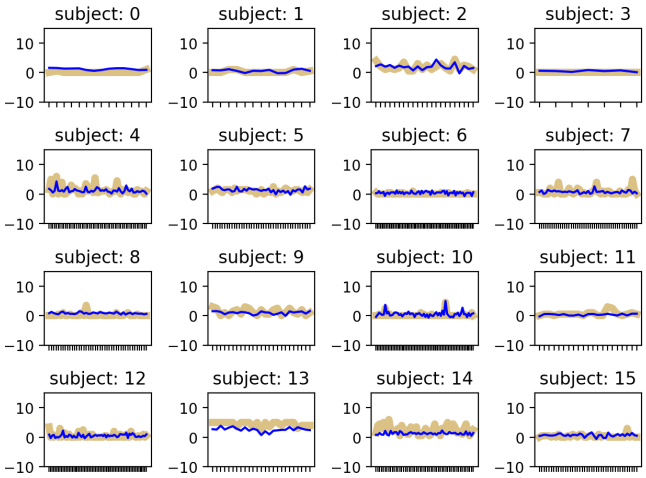

Fig. 8: The LMEM for digiDay_M, separated by subjects and ordered in time. The brown line is the self-reported ground truth. The blue line is the estimated value. These data were generated through **data fitting**.

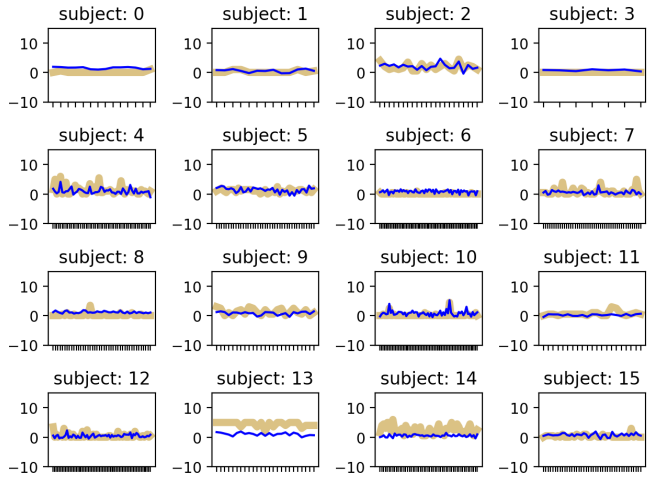

Fig. 9: The LMEM for digiDay_M, separated by subjects and ordered in time. The brown line is the self-reported ground truth. The blue line is the estimated value. These data were generated through **subject independent prediction**.

X. Limitations and Future Work

The presented work is the result of a pilot deployment. Although the dataset itself is quite large, the labels over which to validate the findings are relatively small. This may impact the generalizability of the work. However, we observe that the performance of the model is aligned with our clinical understanding of expression of mood severity for individuals with bipolar disorder. We are continuing to collect data using this pipeline.

The classifiers discussed in this paper are based on perception-of-other labels, yet EMA is self-report. There is an inherent mismatch between these two types of labels [28], [106]). Future work will investigate additional continual learning approaches to enable the automatic prediction of self-reported emotion.

XI. Conclusions

In this paper, we presented a novel emotion-centered data collection in real-world environments for individuals with bipolar disorder. We explored the relationship between EMA (self-report) and passive emotion estimates (perception-of-other), describing how the two emotion measures differ. This included assessment of emotion features that could be extracted from each, focused on the daily mean emotion, inertia of emotion, variability of emotion, and the presence of anomalies across both passive and EMA. We demonstrated how features derived from these disparate emotion measures could be validated in the context of mental health symptom severity. The resulting LMEM models demonstrate the utility of both passive and EMA approaches, particularly as a function of the level of insight that is associated with both manic and depressive symptom severity.

XII. Ethical Considerations

All work presented in this paper has been approved by University of Michigan's IRB. Yet, the use of emotion recognition technology is not without potential harm. In this paper, emotion recognition is presented in the context of understanding variation in mental health symptom severity, which has significant potential benefits surrounding the recognition of early warning signs of illness, thus enabling just-in-time intervention. However, the use of these technologies must be carefully considered. Recent research has focused on identifying potential harms and necessary legal protections as it relates to emotion ethics and privacy [107]–[110]. McStay and Pavliscak have created guidelines for the ethical use of emotional AI, focusing on personal questions (e.g., benefits to the user, respecting human dignity, facilitating meaningful choice), relationship questions, and societal questions [110]. Roemmich and Andalibi investigated how data subjects themselves perceive emotion recognition technologies. They identify potential harms including "the spread of inaccurate health information, inappropriate surveillance, and interventions informed by inaccurate predictions [108]". Grill and Andalibi point towards data subjects' discomfort, concern with agency, and the lack of transparency within the technology [109]. We take these guidelines and suggestions into account, consistently considering the tradeoffs that exist between the benefits of this technology with respect to individual health and the ability to provide new insight into the time course of illness, while mitigating potential negative risks to data subject's health, safety, and autonomy.

XIII. Acknowledgements

This material is based in part upon work supported by the National Science Foundation (NSF IIS-RI 2230172), National Institutes of Health: R34MH100404 and UM1TR004404, Baszucki Group, Prechter Bipolar Research Program, the Tam Foundation, and an Investigators Awards grant program of Precision Health at the University of Michigan. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

References

[1] Dror Ben-Zeev, Rachel Brian, Rui Wang, Weichen Wang, Andrew T Campbell, Min SH Aung, Michael Merrill, Vincent WS Tseng, Tanzeem Choudhury, Marta Hauser, John M. Kane, and Emily A Scherer. Crosscheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric Rehabilitation Journal*, 40(3):266, 2017.

[2] Shefali Kumar, Megan Jones Bell, and Jessie L Juusola. Mobile and traditional cognitive behavioral therapy programs for generalized anxiety disorder: A cost-effectiveness analysis. *PloS One*, 13(1):e0190554, 2018.

[3] Jennifer Nicholas, Andrea S Fogarty, Katherine Boydell, and Helen Christensen. The reviews are in: A qualitative content analysis of consumer perspectives on apps for bipolar disorder. *Journal of Medical Internet Research*, 19(4), 2017.

[4] Michael F Armey, Heather T Schatten, Natasha Haradhvala, and Ivan W Miller. Ecological momentary assessment (ema) of depression-related phenomena. *Current opinion in psychology*, 4:21–25, 2015.

[5] Ellen Frank, Janice Pong, Yashvi Asher, and Claudio N Soares. Smart phone technologies and ecological momentary data: is this the way forward on depression management and research? *Current Opinion in Psychiatry*, 31(1):3–6, 2018.

[6] Hugo Vachon, Wolfgang Viechtbauer, Aki Rintala, and Inez Myin-Germeys. Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *Journal of Medical Internet Research*, 21(12):e14475, 2019.

[7] Frederick K. Goodwin, Kay R. Jamison, and S. Nassir Ghaemi. *Manic-depressive illness: bipolar disorders and recurrent depression*. Oxford University Press, New York, 2nd edition, 2007.

[8] Tuka Al Hanai, Mohammad Ghassemi, and James Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720, Hyderabad, India, 2018.

[9] James R Williamson, Diana Young, Andrew A Nierenberg, James Niemi, Brian S Helfer, and Thomas F Quatieri. Tracking depression severity from audio and video based on speech articulatory coordination. *Computer Speech & Language*, 55:40–56, 2019.

[10] Zhaocheng Huang, Julien Epps, Dale Joachim, and Michael Chen. Depression detection from short utterances via diverse smartphones in natural environmental conditions. pages 3393–3397, Hyderabad, India, 2018.

[11] Brian Stasak and Julien Epps. Differential performance of automatic speech-based depression classification across smartphones. In *Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 171–175. IEEE, 2017.

[12] Nadee Seneviratne, James R Williamson, Adam C Lammert, Thomas F Quatieri, and Carol Espy-Wilson. Extended study on the use of vocal tract variables to quantify neuromotor coordination in depression. In *Interspeech*, pages 4551–4555, Shanghai, China, 2020.

[13] Nadee Seneviratne and Carol Espy-Wilson. Deep learning based generalized models for depression classification. *arXiv preprint arXiv:2011.06739*, 2020.

[14] Zhaocheng Huang, Julien Epps, Dale Joachim, Brian Stasak, James R Williamson, and Thomas F Quatieri. Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated CNNs. pages 4561–4565, Shanghai, China, 2020.

[15] Sadari Jayawardena, Julien Epps, and Eliathamby Ambikairajah. Ordinal logistic regression with partial proportional odds for depression prediction. *IEEE Transactions on Affective Computing*, 2020.

[16] Zhaocheng Huang, Julien Epps, and Dale Joachim. Exploiting vocal tract coordination using dilated CNNs for depression detection in naturalistic environments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6549–6553, Barcelona, Spain, 2020.

[17] Soheil Khorram, Mimansa Jaiswal, John Gideon, Melvin McInnis, and Emily Mower Provost. The PRIORI emotion dataset: Linking mood to emotion detected in-the-wild. In *Interspeech*, Hyderabad, India, 2018.

[18] Soheil Khorram, John Gideon, Melvin McInnis, and Emily Mower Provost. Recognition of depression in bipolar disorder: Leveraging cohort and person-specific knowledge. In *Interspeech*, pages 1215–1219, San Francisco, CA, 2016.

[19] Zahi N Karam, Emily Mower Provost, Satinder Singh, Jennifer Montgomery, Christopher Archer, Gloria Harrington, and Melvin Mcinnis. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4858–4862, Florence, Italy, May 2014.

[20] John Gideon, Emily Mower Provost, and Melvin McInnis. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2359–2363, Shanghai, China, 2016.

[21] M Faurholt-Jepsen, Jonas Busk, M Frost, M Vinberg, EM Christensen, Ole Winther, Jakob Eyvind Bardram, and LV Kessing. Voice analysis as an objective state marker in bipolar disorder. *Translational Psychiatry*, 6(7):e856, 2016.

[22] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, Elvan Çiftçi, Hüseyin Güleç, Albert Ali Salah, and Maja Pantic. AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *ACM Audio/Visual Emotion Challenge and Workshop (AVEC)*, pages 3–13, Seoul, Korea, 2018.

[23] Kun-Yi Huang, Chung-Hsien Wu, Ming-Hsiang Su, and Yu-Ting Kuo. Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model. *IEEE transactions on affective computing*, 11(3):393–404, 2018.

[24] Zakaria Aldeneh, Mimansa Jaiswal, Michael Picheny, Melvin G. McInnis, and Emily Mower Provost. Identifying Mood Episodes Using

Dialogue Features from Clinical Interviews. In *Interspeech*, pages 1926–1930, 10.21437/Interspeech.2019-1878, 2019.

[25] Katie Matton, Melvin G. McInnis, and Emily Mower Provost. Into the Wild: Transitioning from Recognizing Mood in Clinical Interactions to Personal Conversations for Individuals with Bipolar Disorder. In *Interspeech*, pages 1438–1442, Graz, Austria, 2019.

[26] John Gideon, Heather T. Schatten, Melvin G. McInnis, and Emily Mower Provost. Emotion Recognition from Natural Phone Conversations in Individuals with and without Recent Suicidal Ideation. In *Interspeech*, pages 3282–3286, Graz, Austria, 2019.

[27] Brian Stasak, Julien Epps, Nicholas Cummins, and Roland Goecke. An investigation of emotional speech in depression classification. pages 485–489, Singapore, 2016.

[28] Biqiao Zhang and Emily Mower Provost. Automatic recognition of self-reported and perceived emotions. In *Multimodal Behavior Analysis in the Wild*, pages 443–470. Elsevier, 2019.

[29] Sonja Biersack and Vera Kempe. Tracing vocal expression of emotion along the speech chain: Do listeners perceive what speakers feel? In *ISCA Workshop on Plasticity in Speech Perception*, Lisbon, Portugal, 2005.

[30] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335, 2008.

[31] Carlos Busso and Shrikanth S. Narayanan. The expression and perception of emotions: Comparing assessments of self versus others. In *Interspeech*, pages 257–260, Brisbane, Australia, September 2008.

[32] Fabien Ringeval, Andreas Sonderegger, Jens Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, Shanghai, China, 2013.

[33] K.P. Truong, M.A. Neerincx, and D.A. van Leeuwen. Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data. In *Interspeech*, pages 318–321, Brisbane, Australia, Sept. 2008.

[34] Khiet P Truong, David A Van Leeuwen, and Franciska MG De Jong. Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech Communication*, 54(9):1049–1063, 2012.

[35] RH Belmaker. Bipolar disorder. *New England Journal of Medicine*, 351(5):476–486, 2004.

[36] Anastasia K Yocum, Emily Friedman, Holli S Bertram, Peisong Han, and Melvin G McInnis. Comparative mortality risks in two independent bipolar cohorts. *Psychiatry research*, 330:115601, 2023.

[37] June Gruber. Can feeling too good be bad? positive emotion persistence (pep) in bipolar disorder. *Current Directions in Psychological Science*, 20(4):217–221, 2011.

[38] Jiyoung Park, Özlem Ayduk, Lisa O'Donnell, Jinsoo Chun, June Gruber, Masoud Kamali, Melvin G McInnis, Patricia Deldin, and Ethan Kross. Regulating the high: Cognitive and neural processes underlying positive emotion regulation in bipolar i disorder. *Clinical Psychological Science*, 2(6):661–674, 2014.

[39] Klaus R Scherer et al. Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162, 2000.

[40] James A Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145, 2003.

[41] J. Scott, G. Murray, C. Henry, G. Morken, E. Scott, J. Angst, K. R. Merikangas, and I. B. Hickie. Activation in bipolar disorders: A systematic review. *JAMA Psychiatry*, 74(2):189–196, 2017.

[42] Elie Cheniaux, Luis Anunciação, J Landeira-Fernandez, and Antonio Egidio Nardi. Mood or energy/activity symptoms in bipolar mania: which are the most informative? *Trends in Psychiatry and Psychotherapy*, (AheadOfPrint):0–0, 2023.

[43] Association American Psychiatric. *DSM 5*. American Psychiatric Association, 2013.

[44] Mimansa Jaiswal and Emily Mower Provost. Best practices for noise-based augmentation to improve the performance of emotion recognition" in the wild". *arXiv preprint arXiv:2104.08806*, 2021.

[45] James Tavernor, Matthew Perez, and Emily Mower Provost. Episodic Memory For Domain-Adaptable, Robust Speech Emotion Recognition. In *Interspeech*, pages 656–660, Dublin, Ireland, 2023.

[46] Ladislav Mošner, Minhua Wu, Anirudh Raju, Sree Hari Krishnan Parthasarathi, Kenichi Kumatani, Shiva Sundaram, Roland Maas, and Björn Hoffmeister. Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6475–6479, Brighton, UK, 2019.

[47] Masakiyo Fujimoto. Factored Deep Convolutional Neural Networks for Noise Robust Speech Recognition. In *Interspeech*, pages 3837–3841, Stockholm, Sweden, 2017.

[48] Sining Sun, Ching-Feng Yeh, Mari Ostendorf, Mei-Yuh Hwang, and Lei Xie. Training Augmentation with Adversarial Examples for Robust Speech Recognition. In *Proc. Interspeech 2018*, pages 2404–2408, Hyderabad, India, 2018.

[49] Davis Liang, Zhiheng Huang, and Zachary C Lipton. Learning noise-invariant representations for robust speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 56–63, Athens, Greece, 2018.

[50] Hu Hu, Tian Tan, and Yanmin Qian. Generative adversarial networks based data augmentation for noise robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5044–5048, Calgary, Alberta, 2018.

[51] Qiu-Shi Zhu, Jie Zhang, Zi-Qiang Zhang, and Li-Rong Dai. A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1927 – 1939, 2023.

[52] Olivia Flanagan, Amy Chan, Partha Roop, and Frederick Sundram. Using acoustic speech patterns from smartphones to investigate mood disorders: scoping review. *JMIR mHealth and uHealth*, 9(9):e24352, 2021.

[53] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W Schuller. Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition. pages 2327–2331, Shanghai, China, 2020.

[54] Alex Wilf and Emily Mower Provost. Dynamic layer customization for noise robust speech emotion recognition in heterogeneous condition training. *Affective Computing and Intelligent Interaction (ACII)*, 2021.

[55] Youngdo Ahn, Sung Joo Lee, and Jong Won Shin. Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation. *IEEE Signal Processing Letters*, 28:1190–1194, 2021.

[56] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745 – 10759, 2023.

[57] John Gideon, Melvin McInnis, and Emily Mower Provost. Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). *IEEE Transactions on Affective Computing*, 12(4):1055 – 1068, 2021.

[58] Mimansa Jaiswal, Zakaria Aldeneh, and Emily Mower Provost. Controlling for confounders in multimodal emotion classification via adversarial learning. In *ACM International Conference on Multimodal Interaction (ICMI)*, page 174–184, Suzhou, Jiangsu, China, 2019.

[59] Bo-Hao Su and Chi-Chun Lee. A conditional cycle emotion gan for cross corpus speech emotion recognition. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 351–357, Virtual, 2021.

[60] Shaokai Li, Peng Song, and Wenming Zheng. Multi-source discriminant subspace alignment for cross-domain speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2448 – 2460, 2023.

[61] Shaokai Li, Peng Song, Liang Ji, Yun Jin, and Wenming Zheng. A generalized subspace distribution adaptation framework for cross-corpus speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Rhodes Island, Greece, 2023.

[62] Leila Jameel, Lucia Valmaggia, Georgina Barnes, and Matteo Cella. mhealth technology to assess, monitor and treat daily functioning difficulties in people with severe mental illness: A systematic review. *Journal of psychiatric research*, 145:35–49, 2022.

[63] Maria Faurholt-Jepsen, Maj Vinberg, Ellen Margrethe Christensen, Mads Frost, Jakob Bardram, and Lars Vedel Kessing. Daily electronic self-monitoring of subjective and objective symptoms in bipolar disorder - the MONARCA trial protocol (MONitoring, treAtment and pRediCtion of bipolAr disorder episodes): a randomised controlled single-blind trial. *BMJ open*, 3(7):e003353, 2013.

[64] Jakob E Bardram, Mads Frost, Károly Szántó, and Gabriela Marcu. The MONARCA self-assessment system: a persuasive personal monitoring system for bipolar patients. In *ACM SIGHIT International Health Informatics Symposium*, pages 21–30, Miami, FL, 2012. ACM.

[65] Jakob E Bardram, Mads Frost, Károly Szántó, Maria Faurholt-Jepsen, Maj Vinberg, and Lars Vedel Kessing. Designing mobile health technology for bipolar disorder: a field trial of the MONARCA system.

In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2627–2636, Paris, France, 2013.

[66] Dror Ben-Zeev, Rachel M Brian, Geneva Jonathan, Lisa Razzano, Nicole Pashka, Elizabeth Carpenter-Song, Robert E Drake, and Emily A Scherer. Mobile health (mhealth) versus clinic-based group intervention for people with serious mental illness: A randomized controlled trial. *Psychiatric Services*, 69(9):978–985, 2018.

[67] Maria Faurholt-Jepsen, Jonas Busk, Helga órarinsdóttir, Mads Frost, Jakob Eyvind Bardram, Maj Vinberg, and Lars Vedel Kessing. Objective smartphone data as a potential diagnostic marker of bipolar disorder. *Australian & New Zealand Journal of Psychiatry*, 53(2):119–128, 2019.

[68] Kelly Ann Ryan, Pallavi Babu, Rebecca Easter, Erika Saunders, Andy Jinseok Lee, Predrag Klasnja, Lilia Verchinina, Valerie Micol, Brent Doil, Melvin G McInnis, et al. A smartphone app to monitor mood symptoms in bipolar disorder: Development and usability study. *JMIR mental health*, 7(9):e19476, 2020.

[69] Gideon P Dunster, Joel Swendsen, and Kathleen Ries Merikangas. Real-time mobile monitoring of bipolar disorder: a review of evidence and future directions. *Neuropsychopharmacology*, 46(1):197–208, 2021.

[70] Evan H Goulding, Cynthia A Dopke, Rebecca Rossom, Geneva Jonathan, David Mohr, and Mary J Kwasny. Effects of a smartphone-based self-management intervention for individuals with bipolar disorder on relapse, symptom burden, and quality of life: A randomized clinical trial. *JAMA psychiatry*, 80(2):109–118, 2023.

[71] Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal*, 38(3):218, 2015.

[72] Rui Wang, Weichen Wang, Min SH Aung, Dror Ben-Zeev, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A Scherer, and Megan Walsh. Predicting symptom trajectories of schizophrenia using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):110, 2017.

[73] Emily Eisner, Sandra Bucci, Natalie Berry, Richard Emsley, Christine Barrowclough, and Richard James Drake. Feasibility of using a smartphone app to assess early signs, basic symptoms and psychotic symptoms over six months: A preliminary report. *Schizophrenia research*, 208:105–113, 2019.

[74] Joy He-Yueya, Benjamin Buck, Andrew Campbell, Tanzeem Choudhury, John M Kane, Dror Ben-Zeev, and Tim Althoff. Assessing the relationship between routine and schizophrenia symptoms with passively sensed measures of behavioral stability. *NPJ schizophrenia*, 6(1):1–8, 2020.

[75] Benjamin Buck, Kevin A Hallgren, Andrew T Campbell, Tanzeem Choudhury, John M Kane, and Dror Ben-Zeev. mhealth-assisted detection of precursors to relapse in schizophrenia. *Frontiers in Psychiatry*, 12:642200, 2021.

[76] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp)*, pages 3–14, Seattle, WA, 2014.

[77] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T Campbell. Unobtrusive sleep monitoring using smartphones. In *IEEE International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 145–152, Venice, Italy, 2013.

[78] Nicholas D Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. Bewell: A smartphone application to monitor, model and promote wellbeing. In *Int. Conference on Pervasive Computing Technologies for Healthcare*, pages 23–26, Dublin, Ireland, 2011.

[79] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *International Conference on Ubiquitous Computing (Ubicomp)*, pages 385–394, Beijing, China, 2011. ACM.

[80] Hong Lu, Jun Yang, Zhigang Liu, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. The jigsaw continuous sensing engine for mobile phone applications. In *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pages 71–84, Zurich, Switzerland, 2010. ACM.

[81] Sophia Haim, Rui Wang, Sarah E Lord, Lorie Loeb, Xia Zhou, and Andrew T Campbell. The mobile photographic stress meter (MPSM): a new way to measure stress using images. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing and ACM International Symposium on Wearable Computers*, pages 733–742, 2015.

[82] Fanglin Chen, Rui Wang, Xia Zhou, and Andrew T Campbell. My smartphone knows I am hungry. In *Workshop on Physical Analytics*, pages 9–14, 2014.

[83] Ohida Binte Amin, Varun Mishra, and Aarti Sathyanarayana. Investigating social interaction patterns with depression severity across different personality traits using digital phenotyping. In *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–4, Boston, MA, 2023.

[84] R. Lotfian and C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, October-December 2019.

[85] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc.

[86] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019.

[87] Zixing Zhang, Bingwen Wu, and Björn Schuller. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6705–6709, Brighton, UK, 2019.

[88] David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[89] Muqiao Yang, Ian Lane, and Shinji Watanabe. Online Continual Learning of End-to-End Speech Recognition Models. In *Interspeech*, pages 2668–2672, Incheon, Korea, 2022.

[90] Rupal Patel, Kathryn Connaghan, Diana Franco, Erika Edsall, Dory Forgit, Laura Olsen, Lianna Ramage, Emily Tyler, and Scott Russell. "the caterpillar": A novel reading passage for assessment of motor speech disorders. 2013.

[91] J.A. Russell. A circumplex model of affect. *Journal of personality and Social Psychology*, 39(6):1161–1178, 1980.

[92] Tijana Sagorac Gruichich, Juan Camilo David Gomez, Gabriel Zayas-Cabán, Melvin G McInnis, and Amy L Cochran. A digital self-report survey of mood for bipolar disorder. *Bipolar Disorders*, 23:810–820, 2021.

[93] Amy Cochran, Livia Belman-Wells, and Melvin McInnis. Engagement strategies for self-monitoring symptoms of bipolar disorder with mobile and wearable technology: Protocol for a randomized controlled trial. *JMIR research protocols*, 7(5):e130, 2018.

[94] Kaela Van Til, Melvin G McInnis, and Amy Cochran. A comparative study of engagement in mobile and wearable health monitoring for bipolar disorder. *Bipolar Disorders*, 22(2):182–190, 2020.

[95] Rafael Valero Fernández. reliabiliPy: measures of survey domain reliability in Python with explanations and examples. Cronbach´s Alpha and Omegas., January 2022.

[96] Sarah H Sperry and Thomas R Kwapil. Affective dynamics in bipolar spectrum psychopathology: Modeling inertia, reactivity, variability, and instability in daily life. *Journal of Affective Disorders*, 251:195–204, 2019.

[97] Sarah H Sperry, Molly A Walsh, and Thomas R Kwapil. Emotion dynamics concurrently and prospectively predict mood psychopathology. *Journal of Affective Disorders*, 261:67–75, 2020.

[98] Maria Faurholt-Jepsen, Jonas Busk, Jakob Eyvind Bardram, Sharleny Stanislaus, Mads Frost, Ellen Margrethe Christensen, Maj Vinberg, and Lars Vedel Kessing. Mood instability and activity/energy instability in patients with bipolar disorder according to day-to-day smartphone-based data–an exploratory post hoc study. *Journal of Affective Disorders*, 334:83–91, 2023.

[99] Sarah H Sperry and Thomas R Kwapil. Bipolar spectrum psychopathology is associated with altered emotion dynamics across multiple timescales. *Emotion*, 22(4):627–640, 2020.

[100] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, pages 10–25080. Austin, TX, 2010.
[101] Sunny J Dutra, Tessa V West, Emily A Impett, Christopher Oveis, Aleksandr Kogan, Dacher Keltner, and June Gruber. Rose-colored glasses gone too far? mania symptoms predict biased emotion experience and perception in couples. *Motivation and Emotion*, 38:157–165, 2014.
[102] S. Nassir Ghaemi. Feeling and time: The phenomenology of mood disorders, depressive realism, and existential psychotherapy. *Schizophrenia Bulletin*, 33(1):122–130, 2007.
[103] Sheri L Johnson, Anda Gershon, and Vladimir Starov. Is energy a stronger indicator of mood for those with bipolar disorder compared to those without bipolar disorder? *Psychiatry Research*, 230(1):1–4, 2015.
[104] Elie Cheniaux, Rafael de Assis da Silva, Cristina MT Santana, Antonio Egidio Nardi, and Alberto Filgueiras. Mood versus energy/activity symptoms in bipolar disorder: which cluster of hamilton depression rating scale better distinguishes between mania, depression, and euthymia? *Trends in psychiatry and psychotherapy*, 41:401–408, 2020.
[105] Rodrigo Machado-Vieira, David A Luckenbaugh, Elizabeth D Ballard, Ioline D Henter, Mauricio Tohen, Trisha Suppes, and Carlos A Zarate Jr. Increased activity or energy as a primary criterion for the diagnosis of bipolar mania in dsm-5: findings from the step-bd study. *American Journal of Psychiatry*, 174(1):70–76, 2017.
[106] Biqiao Zhang, Georg Essl, and Emily Mower Provost. Automatic recognition of self-reported and perceived emotion: Does joint modeling help? In *International Conference on Multimodal Interaction (ICMI)*, page 217–224, Tokyo, Japan, November 2016.
[107] Luke Stark and Jesse Hoey. The ethics of emotion in artificial intelligence systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 782–793, Virtual, 2021.
[108] Kat Roemmich and Nazanin Andalibi. Data subjects' conceptualizations of and attitudes toward automatic emotion recognition-enabled wellbeing interventions on social media. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34, 2021.
[109] Gabriel Grill and Nazanin Andalibi. Attitudes and folk theories of data subjects on transparency and accuracy in emotion recognition. *Proc. of the ACM on Human-Computer Interaction*, 6(CSCW1):1–35, 2022.
[110] Andrew McStay and Pamela Pavliscak. Emotional artificial intelligence: Guidelines for ethical use. https://drive.google.com/file/d/1frAGcvCY_v25V8ylqgPF2brTK9UVj_5Z/view, 2023.

Emily Mower Provost (M'11, SM'17) is a Professor in Computer Science and Engineering at the University of Michigan. Dr. Mower Provost received her Ph.D. in Electrical Engineering from the University of Southern California (USC), Los Angeles, CA in 2010. She is a Toyota Faculty Scholar (2020) and has been awarded a National Science Foundation CAREER Award (2017), the Oscar Stern Award for Depression Research (2015), a National Science Foundation Graduate Research Fellowship (2004-2007). She is an Associate Editor for IEEE Transactions on Affective Computing and the IEEE Open Journal of Signal Processing. She has also served as Associate Editor for Computer Speech and Language and ACM Transactions on Multimedia. She has received best paper awards or finalist nominations for Interspeech 2008, ACM Multimedia 2014, ICMI 2016, and IEEE Transactions on Affective Computing. Among other organizational duties, she has been Program Chair for ACII (2017, 2021), ICMI (2016, 2018). Her research interests are in human-centered speech and video processing, multimodal interfaces design, and speech-based assistive technology. The goals of her research are motivated by the complexities of the perception and expression of human behavior.

Sarah Sperry is an Assistant Professor in Psychiatry and an Adjunct Assistant Professor in Psychology at the University of Michigan. Dr. Sperry obtained her Ph.D. in Clinical Psychology at the University of Illinois at Urbana-Champaign, completed her clinical internship at the Medical University of South Carolina Charleston Consortium, and conducted postdoctoral work at Vanderbilt University Medical Center. She is the director of the Emotion and Temporal Dynamics (EmoTe) Lab and is one of the key investigators within the Heinz C. Prechter Bipolar Research Program. Her broad mission is to improve early detection, predict illness trajectory, and develop personalized interventions for bipolar spectrum disorders. Within this broader mission, she uses mobile technology (smartphones, wearable devices) and intensive longitudinal modeling to characterize and understand intraindividual variability in real-world contexts with a focus on emotion, sleep and circadian rhythms, substance use, and impulsivity. She is also a licensed clinical psychologist, actively engaged in treating adults with bipolar and related disorders using evidenced-based interventions. She supervises and teaches practicum students, postdoctoral fellows, and other clinical trainees.

James Tavernor is a third-year Computer Science and Engineering Ph.D. Candidate at the University of Michigan. In 2020, he received a Master of Engineering in Joint Honours Mathematics and Computer Science from Imperial College London, London, United Kingdom. His research interests lie in the intersection of domain acoustic speech adaptation and speech emotion recognition. His work for the PRIORI project has involved the development of robust emotion recognition approaches.

Steve Anderau is a Data Manager/App Programmer with the Heinz C. Prechter Bipolar Research Program at Michigan Medicine. He is pursuing his master's degree in Software Engineering with a focus on mobile and cloud computing at the University of Michigan Dearborn. In addition to the PRIORI application management, verification, and validation, his work with the Prechter Bipolar Research Program includes full stack engineering and data management across several research projects.

Anastasia Yocum is the Lead Database Architect at the Prechter Bipolar Research Program. Dr. Yocum is focused on the design and development of the data management with the goal of increasing progress in treating bipolar disorder. She is a co-founder of a bioinformatics CRO/consulting group, A2IDEA. Previously, she has worked as a Research Scientific Facilitator in Precision Health at the University of Michigan, with the goal of understanding the various regulatory and infrastructure requirements for integrating and safely sharing the data related to the electronic health record. She holds degrees from the University of Delaware, Drexel University and finally, a Ph.D. in Pharmacology from the University of Pennsylvania.

Melvin McInnis is the Thomas B and Nancy Upjohn Woodworth Professor of Bipolar Disorder and Depression and the Director of the Heinz C Prechter Bipolar Research Program at the University of Michigan (U-M). Dr. McInnis is an internationally recognized expert in bipolar and depressive disorders. He completed his medical and psychiatric training the University of Iceland and the Maudsley Hospital, Kings College London, and fellowship training in medical and psychiatric genetics at the Johns Hopkins University. He directs a comprehensive clinical consultative program in bipolar disorder at the U-M and is active in community advocacy, outreach, and educational programs. He is a Fellow of the Royal College of Psychiatry, the Royal Society of Medicine, and the American College of Neuropsychopharmacology. In 2018 he was awarded the Scientific Research Prize from the National Alliance for Mental Illness.