Multimodal Mental Health Digital Biomarker Analysis From Remote Interviews Using Facial, Vocal, Linguistic, and Cardiovascular Patterns

Zifan Jiang, Salman Seyedi, Emily Griner, Ahmed Abbasi, *Senior Member, IEEE*, Ali Bahrami Rad, *Member, IEEE*, Hyeokhyen Kwon, Robert O. Cotes, and Gari D. Clifford, *Fellow, IEEE*

Abstract*—*Objective: **Psychiatric evaluation suffers from subjectivity and bias, and is hard to scale due to intensive professional training requirements. In this work, we investigated whether behavioral and physiological signals, extracted from tele-video interviews, differ in individuals with psychiatric disorders.** ***Methods:*** **Temporal variations in facial expression, vocal expression, linguistic expression, and cardiovascular modulation were extracted from simultaneously recorded audio and video of remote interviews. Averages, standard deviations, and Markovian process-derived statistics of these features were computed from 73 subjects. Four binary classification tasks were defined: detecting 1) any clinically-diagnosed psychiatric disorder, 2) major depressive disorder, 3) self-rated depression, and 4) self-rated anxiety. Each modality was evaluated individually and in combination.** ***Results:*** **Statistically significant feature differences were found between psychiatric and control subjects. Correlations were found between features and self-rated depression and anxiety scores. Heart rate dynamics provided the best unimodal performance with areas under the receiver-operator curve (AUROCs) of 0.68–0.75 (depending on the classification task). Combining multiple modalities provided AUROCs of 0.72–0.82.** ***Conclusion:*** **Multimodal features extracted from remote interviews revealed informative characteristics of clinically diagnosed and self-rated mental health status.** ***Significance:*** **The proposed multimodal approach has the potential to facilitate scalable, remote, and low-cost assessment for low-burden automated mental health services.**

***Index Terms*—Telehealth, digital biomarker, multimodal, depression, anxiety, mental health, remote photoplethysmography, computer vision, foundation model, machine learning.**

I. Introduction

THE World Health Organization estimated that 13% of the world population, or close to one billion people worldwide, live with a mental disorder, where most of them do not have access to effective care [1]. In addition to being the second most common cause of years of life lived with disability worldwide [2], this crisis of psychiatric disorders translates to an economic burden of $280 billion every year in the United States alone [3]. To reduce the financial cost and to delay the transition into often chronic or life-long psychiatric conditions, it is critical to gain a better understanding and to provide an objective, fast, and accessible evaluation of those disorders to enable early and effective interventions. However, the present diagnosis and phenotyping of psychiatric disorders fail to fully satisfy this dire need due to its subjectivity and biases, and access to psychiatric care is limited even in high-income countries such as the US [4].

The current clinical practice diagnoses psychiatric disorders such as depression and anxiety disorder using the subjective clinical evaluation of signs and symptoms specified by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [5] or the International Classification of Diseases, 10th revision [6]. These diagnostic criteria often suffer from low inter-rater reliability. In the DSM-5 field trials [7], inter-rater reliability (Cohen's kappa, κ) was just 0.28 for a diagnosis of major depressive disorder (MDD) and 0.20 for general anxiety disorder (GAD). Factors such as differences in training, biases (race, gender, culture), and interview style were the most common explanations

Manuscript received 26 September 2023; revised 7 December 2023; accepted 6 January 2024. Date of publication 10 January 2024; date of current version 7 March 2024. This work was supported in part by the National Institute on Deafness and Other Communication Disorders under Grant 1R21DC021029-01A1, in part by the Emory School of Medicine's Imagine, Innovate, Impact Funds, and in part by the Georgia Clinical & Translational Science Alliance National Institutes of Health under Grant UL1-TR002378. *(Corresponding author: Zifan Jiang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by The Emory University Institutional Review Board and the Grady Research Oversight Committee under Application No. IRB# 00105142.

Zifan Jiang and Gari D. Clifford are with the Department of Biomedical Informatics, Emory School of Medicine, Atlanta, GA 30322 USA, also with the Department of Biomedical Engineering, Emory University and Georgia Institute of Technology, Atlanta, GA 30322 USA, and also with the Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: zifanjiang@gatech.edu; gari@gatech.edu).

Salman Seyedi, Ali Bahrami Rad, and Hyeokhyen Kwon are with the Department of Biomedical Informatics, Emory School of Medicine, Atlanta, GA 30322 USA (e-mail: sseyedi@dbmi.emory.edu; ali.bahrami.rad@dbmi.emory.edu; hyeokhyen.kwon@emory.edu).

Emily Griner and Robert O. Cotes are with the Department of Psychiatry and Behavioral Sciences, Emory School of Medicine, Atlanta, GA 30322 USA (e-mail: emily.lynn.griner@emory.edu; robert.o.cotes@emory.edu).

Ahmed Abbasi is with the Department of IT, Analytics, and Operations, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: aabbasi@nd.edu).

Digital Object Identifier 10.1109/JBHI.2024.3352075

for discrepancies between raters [8], [9]. Self-rated questionnaires such as General Anxiety Disorder-7 (GAD-7) [10], and Patient Health Questionnaire-9 (PHQ-9) [11] are also widely used in practice for initial screening and symptom monitoring purposes. Naturally, these scales are highly subjective as they are self-reported: the symptoms reported tend to be over-reported and more severe than observer ratings and highly depend on the subjective response processes [12].

The rapid development of objective automated digital assessment tools could potentially aid clinicians in diagnosing and evaluating mental illness [13]. Those tools help address the potential bias and inaccuracy of the current diagnosing practices by providing an objective and more quantified measurement of behavioral and physiological symptoms. Research groups have developed tools using various types of data modality, validated in numerous mental health populations, including depression [14], [15], anxiety [16], schizophrenia [17], and posttraumatic stress disorder (PTSD) [18]. Diverse modalities of signals have been investigated, including behavioral signals, such as facial and body movements [14], [15], [19], [20], speech acoustics [21], [22], [23], verbal or written content [24], sleep [25] and activity [18], [26] patterns, as well as physiological signals such as cardiovascular (heart rate [18], [27], electrocardiogram [23], [28]) and neural signals (electroencephalogram [29], [30], functional magnetic resonance imaging [31], [32] and functional near-infrared spectroscopy [33]). The multimodal approach, or the combination of multiple types of signals, has been widely adopted to improve the accuracy and robustness of those automated assessments [34], [35]. For example, [36], [37] combined behavioral signals, including cues from video, audio, and text, while others [18], [38] found the combination and interaction between physiological and behavioral signals useful in evaluating disorders.

While the findings in the above studies were promising, there remain unsolved challenges: data in most of them were collected within a lab-controlled environment and/or with specialized hardware, which prohibits potential future access and might not be able to generalize to actual clinical practice. The increasing use of telemedicine in psychiatry in recent years, which was further accelerated by the COVID-19 pandemic [39], provided a promising approach to improve the access and effect of psychiatric care [40], [41], [42], while at the same time presented an unprecedented opportunity of data collection for objective psychiatric assessments development without the limitation of geographical location and specialized hardware [43]. This begs the question of whether data collected remotely, such as in [44], [45], and in our previous research protocol [13], can provide a comparable level of information as the data collected in a lab-controlled environment.

To address those challenges, we investigated whether each and the combination of behavioral and physiological signals, extracted from audio-visual recordings of remote telehealth interviews, which were collected using heterogeneous generic electronic devices (laptops, tablets, or smartphones), were informative in assessing the multiple facets of psychiatric disorders of control subjects and subjects with mental health conditions (MHC). Specifically, we evaluated the differences in the behavioral and physiological features between different diagnostic groups and studied whether mental health conditions could be accurately assessed using those features. Classification instead of regression tasks were utilized because Mini-International Neuropsychiatric Interview (MINI) [46] was used as the primary diagnosing tool in this study, which resulted in binary categorizations (control vs. MHC).

The main contributions of this work are as follows: (1) We showed that audio-visual recordings of remote interviews collected fully remotely and without device limitation could be used to assess mental health states, with similar performance compared to the performance shown in previous studies where data were collected from lab-controlled environments. (2) We proposed a multimodal machine learning analysis framework, where we extracted both hand-crafted features and self-supervised-learned representations of facial, vocal, linguistic, and remote photoplethysmography (rPPG) patterns using signal processing approaches and state-of-the-art deep learning models, including convolutional neural networks (CNN) and transformer-based [47] foundation models. (3) Using those features and derived temporal dynamics, we compared the performance of features extracted from different modalities, with different models, and the performance of the combined features of multiple modalities, in classifying states of depression, anxiety, and absence of any diagnosed disorder using both self-reported scales (PHQ-9, GAD-7) and clinical diagnoses made by clinicians.

II. Dataset

A. Participants

The overall recruitment protocol can be found in Cotes et al. [13], which was designed to recruit three outpatient groups: 50 schizophrenia patients, 50 unipolar major depressive disorder patients, and 50 individuals with no psychiatric history. Due to the difficulty of recruiting enough in-person schizophrenia subjects during COVID-19, in this work, we focused on analyzing subjects recruited as control and depressed subjects. A total of 84 subjects were recruited as of July 17th, 2023, excluding schizophrenia subjects. The Emory University Institutional Review Board and the Grady Research Oversight Committee granted approval for this study (IRB# 00105142). Interviewees were recruited from Research Match (researchmatch.org), a National Institutes of Health-funded online recruitment strategy designed to connect potential participants to research studies, and through Grady's Behavioral Health Outpatient Clinic utilizing a database of interested research participants. Participants were aged $18-65$ and were native English speakers. For the initial screening, interviewees were recruited for either a control group (no history of mental illness within the past 12 months) or a group currently experiencing depression. All diagnoses and group categorizations were verified and finalized by the overseeing psychiatrist and clinical team after the semi-structured interview.

Two subjects did not meet the inclusion criteria based on the information shared during the interview. Interviews from four subjects were accidentally interrupted or unrecorded due to technical issues with the subjects' devices, and the recorded audio or video files from five subjects were corrupted or led to signal extraction errors in certain modalities (for example, rPPG extraction error due to large percentage of facial occupation due to large yaw angle). Hence, data from 73 subjects were included in the analyses. Table I shows the demographics of those included participants.

TABLE I
DEMOGRAPHICS OF THE SUBJECTS GROUPED BY DIAGNOSES

	Controls	MHC
Number of Subjects	22	51
Age (Years)	42.7 ± 14.0	36.6 ± 13.2
Gender (M/F/NB/NA)	9/13/0/0	10/38/2/1
Race (W/B/A/H/O/NA)	10/7/2/0/2/1	28/10/9/2/2/0
Years of Education†	17.3 ± 4.6	16.7 ± 2.5

± indicates the standard deviation of the measured variable. Subjects with current mental health conditions or a history of diagnosis within 12 months were grouped as "MHC", while the rest were considered "Controls". For gender, "M" = male, "F" = female, "NB" = non-binary, and "NA" = no answer. For race, "W" = white, "B" = Black, "A" = Asian, "H" = Hispanic, "O" = more than one race, and "NA" = no answer. The year of education indicates the number of academic years a person completed in formal programs. High school completion usually corresponds to 12 years of education, whereas college completion usually corresponds to 16 years. † Education levels from two subjects were not recorded, and therefore the last entry is based on 22 Controls and 49 MHC individuals. No significant differences (Mann-Whitney, $p > 0.05$) were found in ages and years of education between Controls and MHC.

B. Interviews and Measurements

The study team created the interview guide and protocol and have components that simulate a psychiatric intake interview [13]. All interviews were conducted remotely via Zoom's secure, encrypted, HIPAA-compliant telehealth platform. Both Video and Audio were recorded. The remote interview was divided into three parts: 1) A semi-structured interview composed of a series of open-ended questions, a thematic apperception test [48], phonetic fluency test [49], and semantic fluency test [50], 2) a sociodemographic section, and 3) clinical assessments which included the MINI 6.0 [46], McGill Quality of Life Questionnaire [51], General Anxiety Disorder-7 [10], and Patient Health Questionnaire-9 [11].

C. Categorization

Subjects were categorized into four different two-class categorizations based on self-rated scales or clinicians' diagnoses to evaluate feature performances in classifying categorizations generated from under different assessment procedures. The mental health assessment task was formulated as classification tasks to align with the clinical practices and mental health screening paradigms.

1) The first and primary categorization is control (n = 22) vs. subjects with mental health conditions (MHC, n = 51) based on diagnoses made using MINI. The characteristics of the two groups can be found in Table I. The latter included subjects diagnosed with any mental health condition currently or a history of diagnosis within 12 months, including disorders like MDD, comorbid or primary GAD, PTSD, panic disorder, social anxiety, agoraphobia, psychotic disorders, manic illnesses, personality disorders, and obsessive-compulsive disorder. The control group included the remaining subjects, who could have mild suicidality, mild agoraphobia, mild substance abuse and dependence, or a remote history (not in the previous 12 months) of MDD and not currently on an antidepressant medication.
   The following three categorizations only included a subset of subjects due to inclusion/exclusion criteria and missing self-rating results. The self-reported scales were also dichotomized to align with the primary categorization for easier performance comparison and cross-categorization analysis.
2) The second categorization is non-MDD-control (n = 18) vs. MDD (n = 38, past or current). Since both groups in the first categorization are heterogeneous, we used this categorization to assess further whether differences could be found between controls and subjects with past or current MDD, which were diagnosed using MINI and supported by self-reported PHQ-9 scores. In this case, we defined non-MDD-control as subjects with no lifetime history of MDD or other mental health conditions (but could have mild suicidality, mild agoraphobia, mild substance abuse and dependence), while the MDD subjects have primary diagnoses of MDD but could include comorbid GAD, PTSD, panic disorder, social anxiety, agoraphobia, and substance use disorder.
3) The third categorization is moderately depressed (PHQ-9 scores > 10, n = 24) vs. rest (PHQ-9 scores $<= 10$, n = 43). PHQ-9 scores were not reported for six subjects, resulting in 67 subjects in this categorization. To evaluate performance in classifying the severity of self-rated depression symptoms, we used a PHQ-9 score-based categorization and adopted a cutoff of 10, which indicates moderate depression [11].
4) The fourth categorization is moderate anxiety (GAD-7 scores > 10, n = 16) vs. rest (GAD-7 scores $<= 10$, n = 49). GAD-7 scores were not reported for 8 subjects, resulting in 65 subjects in this categorization. Similar to the third categorization, we used a GAD-7 score-based categorization and adopted a cutoff of 10, which indicates moderate anxiety and a reasonable cut for identifying cases of GAD [10], to evaluate performance in classifying the severity of self-rated anxiety symptoms.

III. METHODS

A. Multimodal Feature Extraction

Fig. 1 shows the proposed multimodal analysis framework that extracts visual, vocal, language, and rPPG time series signals at the frame or segment level, summarizes those time series with statistical and temporal dynamic features at the subject level (except for text embedding from the large language model, where the model directly generated subject-level embedding), and evaluates the performance of these features in clinical diagnoses or self-rated severity classification tasks described in Section II-C.

1) Facial Expressions and Visual Patterns: We followed the CNN-based facial expression analysis framework we proposed in our previous work [14], [57]. For each frame of the recordings sample at 1 Hz (one frame per second), the face of the participant is detected with RetinaFace [58] using a ResNet-50 [59] backbone network trained on the "WIDER" face dataset [60]. The face detector achieved an accuracy of 95.5% on the "Easy" validation set in WIDER face dataset, where the faces were already much more difficult to detect than the faces in our use case. The segmented face was fed into another CNN with VGG19 [61] structure, which was trained on the "AffectNet" dataset [62], to estimate facial emotion probabilities of seven

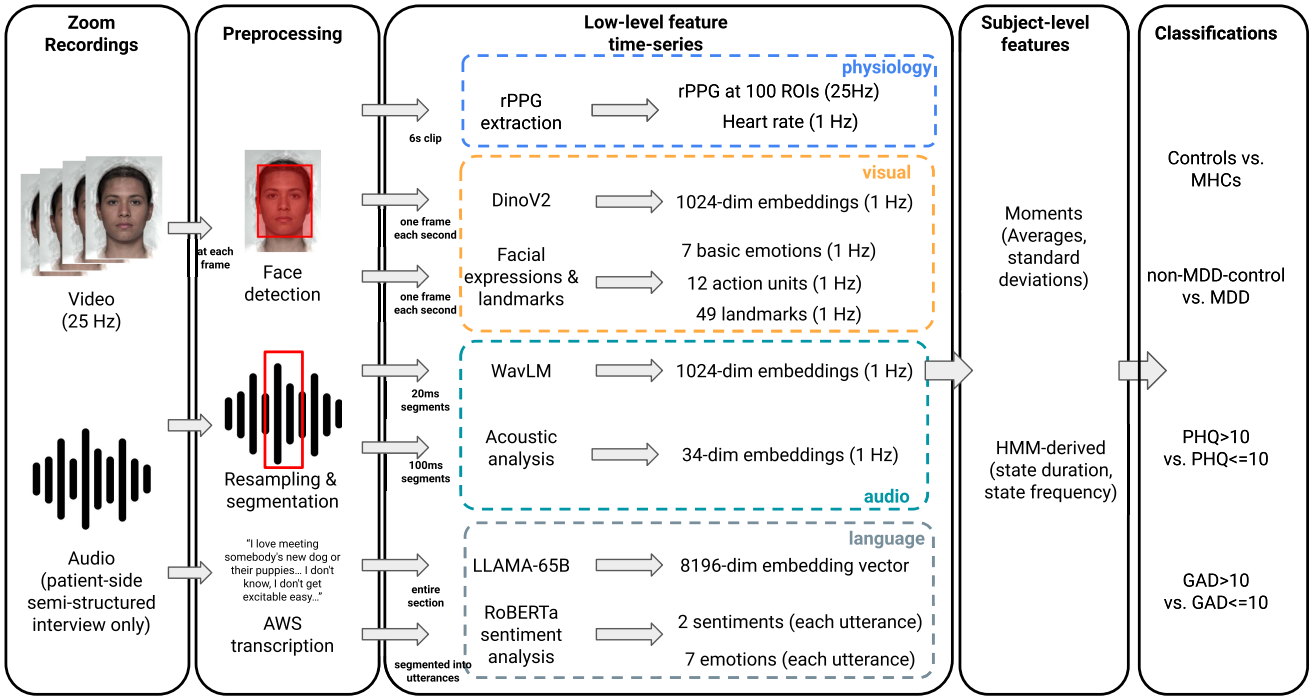

Fig. 1. Overview of the processing pipeline. Color-dashed boxes denote features from different modalities, including physiological, visual, audio, and language features. The audiovisual recordings were first preprocessed with face detection and segmentation for each frame, automatic transcription of the patient-side audio, and audio resampling and segmentation. Then, low-level and subject-level features from various modalities were extracted and used for the classification of mental health conditions. Abbreviations are as follows: "AWS" denotes the Amazon web services, "rPPG" denotes the remote photoplethysmogram extracted from the face [52], "DinoV2" [53], "WavLM" [54], "LLAMA" [55] and "RoBERTa"[56] are the foundation models for each modality; "HMM" denotes the hidden Markov model; "MHC" refers to subjects with mental health conditions; "MDD" refers to subjects with major depressive disorder; "PHQ" and "GAD" refers to the Patient Health Questionnaire-9 and General Anxiety Disorder-7 scores, respectively. Please refer to Section II-C for more details.

categories, including being neutral, happy, sad, surprised, fearful, disgusted, and angry. AffectNet dataset and Radboud Faces Database (RaFD) [63] was used to test this facial emotion classifier. An accuracy of 63.3% was achived in the AffectNet evaluation set and an accuracy of 90.1% was achived in RaFD.

To include facial behaviors less affected by cultural differences, we adopted JAA-Net [64] to recognize 49 facial landmarks and 12 facial action units [65] (AUs, or the individual components of facial muscle movement) expressed in the frame. JAA-Net is a deep learning model that combines CNN and adaptive attention module, and it achieved an average AU detection accuracy of 78.6% (including AU1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, 24) and face alignment mean error of 3.8% inter-ocular distance on BP4D dataset [66] with three-fold cross-validation.

In addition to manually-defined facial expression signals, including facial emotions, AUs, and facial landmark movements, a self-supervised large vision foundation model named "DINOv2" [53] was also used to extract general visual embedding of the segmented facial area. While video foundation models have better performance in short-video clips, the image foundation model was used because the average length of the video recorded in this study was significantly longer (one hour vs. a few seconds). DINOv2 is a vision transformer (ViT) [67] with one billion parameters trained on 1.2B unique images that achieved decent performance on video classification tasks with linear evaluation, including an accuracy of 90.5% on "UCF-101" dataset [68]. A 1024-dimensional visual embedding was generated from frames sampled at 1 Hz using the "ViT-L/14" [67] model.

2) Language Sentiments and Representations: The patient-side audio files were transcribed into texts using Amazon Transcribe on HIPAA-compliant Amazon web services (AWS) at Emory, following the protocol detailed in our previous study [69]. Similar to the audio analysis, only patient-side transcripts during the semi-structured interview section were used to avoid using subjects' answers to sociodemographic or clinical assessment questions.

We have previously found different word use patterns in subjects with and without MDD using the linguistic inquiry and word count (version LIWC-22) dictionary [70]. Here large language models (LLMs) were used to identify the sentiments and extract general representations to better understand the subjects' linguistic patterns. More specifically, three LLMs were used: (1) At the utterance level, a distilled RoBERTa model [56], [71] finetuned on 80% of 20 k emotional texts (the rest 20% was used as the test set with an average accuracy of 66%) was used to recognize one of seven emotions including neutral, happiness, sadness, surprise, fear, disgust, and anger. (2) Also at the utterance level, another RoBERTa-based model finetuned on 15 diverse review datasets with a leave-one-dataset-out accuracy of 93.2% [72] was used to recognize positive or negative sentiment. Such fine-tuned utterance-level deep learning models have been found to generate effective representations in related contexts such as anxiety [73]. (3) LLAMA-65B [55], one of the state-of-the-art open-sourced decoder-only transformer models with 65 billion parameters which were trained on over one trillion tokens of texts, was used to generate an 8196-dimensional text embedding for the entire transcripts during the semi-structured interview.

3) Vocal Features and Representations: Both manually defined acoustic features and general audio representations were extracted from audio files. Only patient-side audio during the semi-structured interview section was used to avoid the potential information leak directly from subjects' answers to sociodemographic or clinical assessment questions in MINI or in self-rated questionnaires described in Section II-B.

For manually defined features, *PyAudioAnalysis* [74] package was used to extract acoustic features at each 100 ms window with 50% overlap, including zero crossing rate, energy, entropy of energy, spectral centroid/spread/entropy/flux/rolloff, Mel frequency cepstral coefficients, and 12 chroma vector and corresponding standard deviations. WavLM [54], which is a self-supervised audio foundation model with 316 M parameters ("WavLM Large") trained on 94 k hours of audio, was used to extract general audio representations. It has shown state-of-the-art performance in the universal speech representation benchmark [75]. Recorded audio files were first resampled to 16 k Hz and then segmented into non-overlapping 20 ms segments following [54]. A 1024-dimensional audio embedding was generated for each 20 ms segment using WavLM.

4) Remote PPG Cardiovascular Features: Remote PPG signals were extracted from the video recordings using the *pyVHR* package [52], [76]. The facial skin areas were recognized in each frame using a CNN, 100 regions of interests (ROIs) were sampled, and the pixel values were averaged across the pixels in each ROI for each RGB channel, respectively. Then, an unsupervised method, named orthogonal matrix image transformation [77], was used to transform RGB values in one ROI to an estimated 25 Hz rPPG signal based on QR decomposition. The power spectral density of rPPGs at each ROI was computed in six seconds windows sliding every second, and the medians of the inverse of peak frequency ($60/$peak frequency) were used to estimate heart rates at every second.

Lastly, the averaged estimated rPPGs at each ROI were used to extract cardiovascular dynamic features using *PhysioNet Cardiovascular Signal Toolbox* [78] with a 300 s window and a 30 s sliding window. The cardiovascular dynamic features included time and frequency domain heart rate variability, acceleration and deceleration capacity, entropy measures, and heart rate turbulence measures. Highly tolerant rejecting thresholds were set to avoid rejecting high percentage of data, including setting lowest tolerable mean signal quality index (as defined in [78]) to be 0.1, allowing certain R-R intervals to be longer than ten seconds, allowing two neighboring R-R intervals to have a length difference of more than one second, and allowing a 30 seconds gap at the beginning of the PPG signals.

B. Subject-Level Features and Temporal Analyses

Due to the high dimensionality of the low-level features and the limited number of subjects, only two simple statistics of the time series extracted above were used as subject-level features to avoid potential overfitting as explored in our previous work [14]. Both average and standard deviations over time were calculated for lower-dimensional (< 100) time series, including time series of facial expressions (facial emotions, AUs, and facial landmark locations sampled at 1 Hz), acoustic features (sampled at 20 Hz), language sentiments (sampled at each utterance), and estimated heart rates (sampled at 1 Hz). Only averages were calculated for higher-dimensional (> 100) time series, including time series of WavLM audio embedding and DINOv2 visual embedding. LLAMA-65B embedding of the entire semi-interviews was directly used as subject-level features.

In addition to nonparametric statistics, hidden Markov models (HMM) were used to model the dynamics of the low dimensional time series, and statistics (duration and frequency of inferred states) of the unsupervisely learned HMMs were used as subject-level features. An HMM with a Gaussian observation model and four states was learned for each modality separately using *SSM* package [79]. The number of states was selected because it represents the smallest number of states needed to model known different states: asymptomatic, symptomatic, uncertain, and padding states. It is worth noting that the states learned from the data do not directly correspond to those four states, nor do we aim to directly interpret those learned states. Instead, we used the downstream analysis of the duration and frequency of the states as an approximate modeling of the dynamics of the time series.

Each time series of one modality from one subject k, X_k, was considered as one noisy observation, where it is padded with zeros to the maximum temporal length $T_{\max}$ found from X_1 to $X_N (N = 73)$. i.e., X_k is a $T_k \times d$ with a feature dimension of d and a temporal length of T_k was padded $(T_{\max} - T_k) \times d$ zeros at the end, so all X_k has the same shape of $T_{\max} \times d$. The modality-specific HMM was then fitted on X, and the most likely hidden states Z_k with the shape of $T_{\max} \times 4$ were inferred for each sequence X_k. Lastly, the time steps spent and the frequency (non-neighboring occurrences) of all four states were calculated for each subject and used as subject-level dynamic features.

C. Classification Analyses

We evaluated features generated from the above-described processes in four two-class classification tasks described in Section II-C. Classification performances were measured by the average area under the receiver operating characteristic (AUROC) and the average accuracy in 100 repeated five-fold cross-validations. In each repetition, subjects were randomly split into five approximately equally sized folds. A cross-validation was performed on those folds, where in each one of the five validations, four folds were used for training and hyper-parameter tuning and one fold left was held out for testing.

1) Demographic Variables: Demographic variables, including one-hot-encoded race, one-hot-encoded gender, age, and years of education, were combined into a demographic feature vector for each subject and also evaluated as a benchmark in unimodal classification. However, demographic features were not considered in the multimodal classification.

2) Unimodal Evaluation: For each type of feature (as shown on each row in Table II extracted from different modalities, statistics (averages and standard deviations) and HMM-derived features were evaluated separately using logistic regression (LR) with $l2$ regularization or a gradient boosting decision tree (GBDT) classifier, depending on the dimensionality of the features, where LR was used for features with fewer than 100 dimensions. For GBDT, a default of 100 base decision tree estimators and a maximum depth of two were set across all types of features.

3) Multimodal Fusion: Both early and late fusion of different modalities were considered. For early fusion, features from all

TABLE II
CLASSIFICATION PERFORMANCE OF CLINICAL DIAGNOSES AND SELF-RATED DEPRESSION/ANXIETY SEVERITY

Feature type	Metric	1. Control vs. MHC	2. Non-MDD-Control vs. MDD	3. PHQ-9 $>$ 10?	4. GAD-7 $>$ 10?
1. Demographic variables	AUROC	0.54 ± 0.04	0.54 ± 0.04	0.61 ± 0.03	0.57 ± 0.05
	Accuracy	0.56 ± 0.04	0.57 ± 0.04	0.58 ± 0.04	0.60 ± 0.04
2. Facial emotions + AUs					
2.1 Avgs and stds	AUROC	random	random	0.56 ± 0.06	0.55 ± 0.04
	Accuracy	random	random	0.60 ± 0.05	0.62 ± 0.04
2.2 HMM features	AUROC	0.65 ± 0.03	0.66 ± 0.04	0.61 ± 0.04	$\mathbf{0.68 \pm 0.05}$
	Accuracy	0.64 ± 0.03	0.66 ± 0.04	0.60 ± 0.04	$\mathbf{0.67 \pm 0.03}$
3. DINOv2 avgs and stds	Both	random	random	random	random
4. Language sentiment					
4.1 Avgs and stds	AUROC	0.69 ± 0.03	0.66 ± 0.04	0.64 ± 0.05	0.63 ± 0.04
	Accuracy	0.67 ± 0.04	0.68 ± 0.04	0.67 ± 0.05	0.64 ± 0.04
4.2 HMM features	AUROC	0.62 ± 0.03	0.64 ± 0.04	random	0.65 ± 0.05
	Accuracy	0.65 ± 0.03	0.60 ± 0.03	random	0.73 ± 0.04
5. LLAMA-65B	AUROC	0.64 ± 0.07	0.53 ± 0.08	0.68 ± 0.04	0.64 ± 0.05
	Accuracy	0.68 ± 0.05	0.59 ± 0.07	0.68 ± 0.03	0.70 ± 0.04
6. WavLM avgs and stds	AUROC	random	0.58 ± 0.05	0.60 ± 0.06	0.59 ± 0.02
	Accuracy	random	0.61 ± 0.05	0.64 ± 0.05	0.71 ± 0.02
7. Vocal acoustics					
7.1 Avgs and stds	AUROC	random	random	0.68 ± 0.05	random
	Accuracy	random	random	0.67 ± 0.05	random
7.2 HMM features	AUROC	0.57 ± 0.05	0.51 ± 0.06	0.51 ± 0.05	0.53 ± 0.05
	Accuracy	0.59 ± 0.04	0.53 ± 0.05	0.53 ± 0.05	0.60 ± 0.04
8. rPPG					
8.1 Cardiovascular features	AUROC	random	0.55 ± 0.07	0.65 ± 0.04	0.56 ± 0.05
	Accuracy	random	0.61 ± 0.05	0.60 ± 0.04	0.59 ± 0.04
8.2 HMM features	AUROC	$\mathbf{0.72 \pm 0.05}$ †	$\mathbf{0.73 \pm 0.05}$ †	$\mathbf{0.75 \pm 0.03}$ †	$\mathbf{0.68 \pm 0.04}$
	Accuracy	$\mathbf{0.76 \pm 0.04}$ †	$\mathbf{0.73 \pm 0.05}$ †	$\mathbf{0.71 \pm 0.03}$ †	$\mathbf{0.67 \pm 0.03}$
9. Fusion (row 2-8)					
9.1 Feature concatenation	AUROC	0.63 ± 0.07	random	0.59 ± 0.05	0.53 ± 0.05
	Accuract	0.68 ± 0.05	random	0.61 ± 0.04	0.62 ± 0.04
9.2 Majority vote	AUROC	0.70 ± 0.05	0.68 ± 0.07	0.75 ± 0.05	0.71 ± 0.05
	Accuracy	0.73 ± 0.03	0.71 ± 0.04	0.71 ± 0.04	0.76 ± 0.03
9.3 Selected vote	AUROC	$\mathbf{0.82 \pm 0.04}$ ‡	$\mathbf{0.77 \pm 0.02}$ ‡	$\mathbf{0.82 \pm 0.04}$ ‡	$\mathbf{0.72 \pm 0.04}$ ‡
	Accuracy	$\mathbf{0.75 \pm 0.03}$ ‡	$\mathbf{0.76 \pm 0.01}$ ‡	$\mathbf{0.74 \pm 0.04}$ ‡	$\mathbf{0.75 \pm 0.03}$ ‡

Each column shows the performance of two-class classification using one of the four categorizations defined in Section. II-C in the same order. The average and the standard deviation of AUROCs and accuracies (in brackets) from a hundred randomly split five-fold cross-validations are reported. The term *avgs* denotes averages, and *stds* denotes standard deviations. "random" indicates that the classifier performed no significantly better (McNemar's test, $p > 0.05$) than random guessing (AUROC=0.5). The best classification performance in each task (column) achieved by a single modality was shown in bold text, while the second best was underlined. Multiple metrics were underlined or marked bold when no statistical significance (McNemar's test, $p > 0.05$) between classifiers was found. The best classification performance in each task (column) achieved by multimodal fusion was shown in bold text. "†" indicates significantly better performance (McNemar's test, $p < 0.05$) was achieved with the indicated feature type in this classification task (each column) compared to other unimodal features, where "‡" indicates significantly better performance (McNemar's test, $p < 0.05$) was achieved with multimodal voting compared to using any unimodal features.

modalities were concatenated into a single feature vector as the input to a GBDT classifier. For late fusion, the majority vote of each unimodal classifier was used as the multimodal classification results. To avoid noise from classifiers without classification power, we also compared the majority voting results from classifiers that showed non-random (defined as $AUROC > 0.5$) performance in the validation set (a 20% subset within the training fold). The non-random classifiers were re-trained with all the data in the training fold before being used for testing.

D. Statistical Analyses

We used statistical tests to assess the differences in the probability distributions of features between different groups of subjects (such as groups described in Section II-C and demographic groups) and the differences in performance resulting from different features. Mann-Whitney rank tests were applied between features or characteristics of different subject groups to determine whether significant differences exist between the two groups. McNemar's test was used to test the classification disagreement between pairs of classification settings. Wald Test was used to determine if a significant correlation was found between two variables. Statistical significance was assumed at a level of $p < 0.05$ for all tests.

IV. RESULTS

A. Unimodal Feature Patterns Across Groups

Here we performed a selected array of analyses of the clinically relevant patterns found in different modalities in different groups of subjects, providing additional objective evidence to previous clinical observations.

1) Blunted Visual Affect and Increased Sadness in Language: While "blunted affect" was mostly in the context of a negative symptom of schizophrenia, it has been widely reported in other mental disorders like MDD [80], [81], [82] and other non-psychotic disorders [83]. Measured by the sum of average AU intensities over the interview, we found that non-medicated subjects with current MDD had lower facial expressivity compared to non-MDD controls (Mann-Whitney, $p = 0.04$), and subjects with mental health conditions also had lower facial expressivity compared to controls (Mann-Whitney, $p = 0.03$). However, no differences in facial expressivity were found between subjects with past MDD and non-MDD controls, and no statistically significant linear correlations were found between facial expressivity and self-rated PHQ-9 or GAD-7 scores.

Through language sentiment analysis, neither was verbally blunted affect found in the MDD or MHC groups nor language expressivity correlate with self-rated scores. However, the average sadness level expressed in language was found

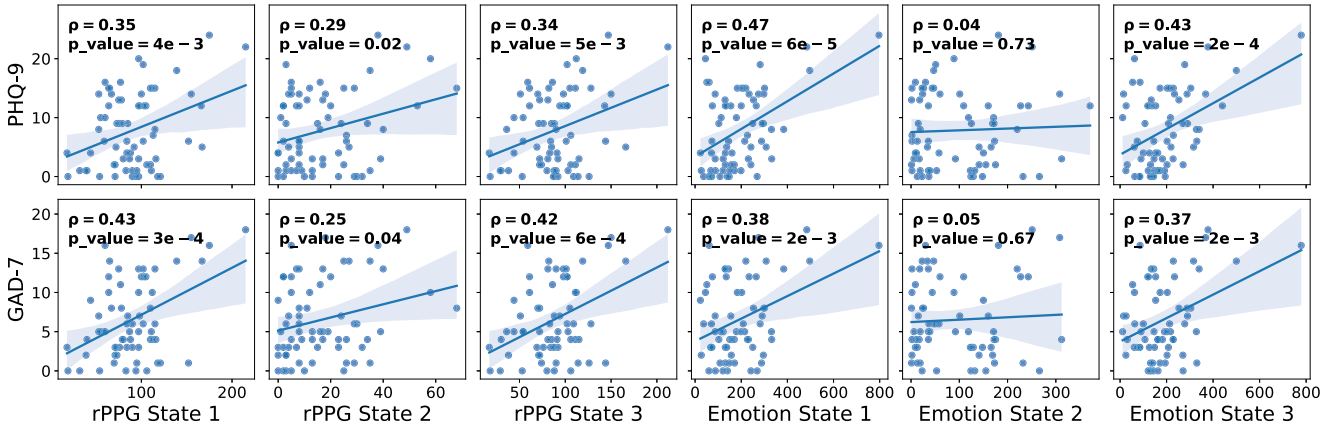

Fig. 2. PHQ-9 and GAD-7 scores vs. rPPG and facial expression HMM state frequencies each subfigure shows the scatter plot between self-rated scores and the frequency of a learned HMM state, along with a linear regression model fit with the 95% confidence interval. The top row shows how those learned state correlate with PHQ-9 scores and the bottom row shows how they correlate with GAD-7 scores. Texts in each subfigure denote the Pearson correlation coefficient (ρ) and the p-value using the Wald test.

to be higher in MDD groups compared to non-MDD-controls (Mann-Whitney, $p = 0.02$) and was positively correlated with PHQ-9 (Wald test, $\rho = 0.31, p = 0.01$) and GAD-7 (Wald test, $\rho = 0.37, p = 0.002$) scores. In comparison, the sadness level expressed visually did not increase in MDD groups.

2) Increased Acoustic Spectral Flux: The average spectral flux, defined as the squared difference between the normalized magnitudes of the spectra of the two successive frames averaged across the semi-structured interviews, was found to be positively correlated with PHQ-9 (Wald test, $\rho = 0.26, p = 0.03$) and GAD-7 (Wald test, $\rho = 0.25, p = 0.04$) scores, indicating a faster change of acoustic tones in subjects with more severe depression and anxiety symptoms.

3) Increased Complexity in Heartbeat Intervals: No significant alternation of average heart rate or standard deviation of heart rate during the interview was found between groups. The complexity of the heartbeat time series, measured by the area under the multiscale entropy curve, was significantly higher in non-medicated MDD groups compared to non-MDD-controls (Mann-Whitney, $p = 0.01$), consistent with previous findings using electrocardiogram [84], [85].

4) Effect of Medication: Compared to non-medicated MDD subjects, medicated MDD subjects showed a higher level of facial expressivity (Mann-Whitney, $p = 0.05$) and sadness (Mann-Whitney, $p = 0.04$), while only non-medicated subjects with current MDD showed a higher level of sadness through language compared to medicated subjects with current MDD (Mann-Whitney, $p = 0.04$). In addition, decreased heartbeat interval complexity (Mann-Whitney, $p = 0.02$) and increased standard deviation of heart rate (Mann-Whitney, $p = 0.02$) were observed with medication in subjects with past and current MDD compared to non-medicated MDD subjects, while the average heart rate remained similar between both groups.

B. Dynamics Inferred From HMM State Duration and Frequency

Dynamic features, including inferred HMM state duration and frequency, were found to be the most useful features in classification tasks, as shown in Table II, especially for facial expressions and rPPG modalities. Significant linear correlations were found between these dynamic features and PHQ-9/GAD-7 scores.

Fig. 2 shows the correlation plots between the frequency of the states in emotion and heart rate time series. The padding states (described in Section III-B) from rPPG and facial expression HMMs were omitted as they would only present once (frequency = 1) as the padding in the end. Statistically significant positive correlations were found between all non-padding state frequency and self-rated scores except emotion state 2, indicating a higher switching rate between hidden states may be related to more severe depression and anxiety symptoms.

C. Classification Performance

Table II shows the classification performance of both clinically diagnosed and self-rated mental health disorders using static and dynamic features from vision, audio, language, and physiology. Each column shows the performance of two-class classification using one of the four categorizations defined in Section II-C in the same order. The best-performing features achieved an AUROC of 0.68 to 0.75 in unimodal classification tasks, while the selected majority voting described in Section III-C3 achieved an AUROC of 0.82 in detecting current or recent (last 12-month) mental disorders, an AUROC of 0.77 in detecting past and current MDD, an AUROC of 0.82 in detecting PHQ-9 based moderate depression, and an AUROC of 0.72 in detecting GAD-7 based moderate anxiety disorder. Late fusion using selected majority voting (row "9.3") outperformed early fusion with the direct concatenation of features (McNemar's test, $p \ll 0.01$) due to the extremely high dimensionality of the concatenated features.

While demographic variables achieved higher than random performance in all four tasks, we found they were not strong predictors of mental health disorders compared to the proposed features, as shown in row "1" in Table II.

A similar level of performance in MDD vs. healthy control classification was achieved compared to results reported in existing studies using in-lab data collection processes. For example, an AUROC of 0.68 was achieved using the facial and speech emotions in our previous in-lab study [14]. Other researchers, such as Schultebraucks et al. [86] achieved an

AUROC of 0.86 combining facial action units, acoustic and language features in another in-lab study. In-lab studies [87], [88] using heart rate variability features also achieved AUROCs of 0.74-0.82, which showed a similar range of performance as with our reported AUROCs when detecting self-reported and clinical MDD, achieved by the rPPG-based method proposed in this study. Please note those performance metrics cannot be directly compared as different subjects, data collection hardware and processes, categorization criteria, and evaluation methods were adopted.

1) Moments and Dynamics of Facial Expressions Revealed Mental States but General Visual Patterns Did Not: While we also extracted facial landmarks as described in Section III-A1, we found adding static statistics of facial landmarks or including facial landmarks in the HMM modeling deteriorated the performance. Row "2" in Table II shows the performance using just the statistics of facial emotions and AUs. Interestingly, the average and standard deviation of facial expressions failed to classify clinical diagnoses but successes in classifying self-rated depression and anxiety. In comparison, the temporal properties derived from HMM resulted in significantly better (McNemar's test, $p \ll 0.01$) classification performance, except for self-rated depression detection. Lastly, using the temporal dynamics of facial expressions achieved the best performance in self-rated anxiety in all modalities.

In comparison, visual embedding generated from DINOv2 failed to generalize to this specialized dataset and did not achieve non-random classification in any of the tasks.

2) Language Sentiments Beat General Language Representation in Small and Specialized Dataset: Compared to other modalities, language features were extracted at a lower sampling rate (at each utterance or the entire semi-structured interview), while LLMs were able to abstract the texts into much shorter sequences of features or even into a single vector when LLAMA-65B was used. The average and standard deviation of the language sentiments achieved the best performances compared to static features of other modalities. Using HMM to model the sentiment dynamics did not improve performance, as shown in all other modalities (comparing rows "4.2" and "4.1"). These results showed that part of the dynamics expressed through the words was already captured by LLM and abstracted into utterance sentiments, and the sentiment dynamics over multiple utterances might not be as important.

Additionally, while using LLAMA-65B embedding showed decent performance compared to other non-language modalities, using language sentiments achieved similar or better results in all tasks. This showed that general language representation might not be as useful as disorder-related sentiment analysis, especially in smaller and highly-specialized datasets, as demonstrated in this study, and suggested in related work on text-based depression and personality detection [89], [90].

3) Vocal Features Were Under-Performing Compared to Other Modalities: While many previous studies [21], [22], [91] have shown that vocal features are useful in detecting depression and anxiety disorder, in this study, other modalities outperformed both spectral/entropy-based acoustic features and general speech representation from WavLM except in self-rated depression detection.

4) HMM Modeled Dynamics Were More Informative Compared to Cardiovascular Features for Highly Noisy rPPG Signals: As shown in row "8.1" in Table II, using cardiovascular features yielded inferior performance compared to other modalities. The key reason is the estimated rPPG signals were highly noisy at each ROI or after averaging across all ROIs, which led to errors (such as peak detection error) in downstream cardiovascular feature calculations. On average, 25.8% of the estimated rPPGs were not used for downstream analyses even with highly tolerant rejecting thresholds as described in Section III-A4. Using HMM-derived features from modeling heart rate time series resulted in the best or second-best performance in all four tasks among all unimodal approaches, reaching AUROCs from 0.68 to 0.75.

5) False Positives in the View of Self-Reported Depression Were Not Necessarily False in the Clinical View: When looking at the false positives (false classification as depression when evaluating with self-reported labels) of the best-performing multimodal classifier, we found that 85% of those cases were actually correctly classified in the view of the clinicians. I.e., that 85% of cases had a current/past MDD or other comorbid mental health condition clinically and were correctly captured by the classifier trained with self-reported PHQ-9-based labels. Although it requires further investigations, this showcased that the model trained with self-reported labels can be helpful for clinical assessments.

V. Discussion and Conclusion

In this work, we performed a detailed multimodal analysis on 73 subjects using remotely-recorded telehealth interviews and showed that the facial, vocal, linguistic, and cardiovascular features extracted from these audiovisual recordings could reveal informative characteristics of both clinically diagnosed and self-rated mental health status. The results provided early evidence of the usefulness of multimodal digital biomarkers extracted from low-cost and non-lab-controlled data with minimal hardware limitations. Comparisons were made between different modalities and between features derived from the latest transformer-based foundation models and more defined features derived from traditional methods, offering insights on which modalities and methods might be most suited for automated remote mental health assessments.

A. Performance of Different Modalities

When comparing the classification performance using features extracted from different modalities, the overall physiological characteristics outperformed other manually-defined or data-driven behavioral characteristics. Although the heart rates were estimated indirectly from light changes on the face, heart rate dynamics were highly relevant in classifying self-rated and clinician-diagnosed disorders. While it is not surprising to find associations between cardiovascular dynamics and psychiatric disorders, as shown in previous studies of neurobiological mechanisms [92] and statistical analyses [88], [93], the results raised questions on the behavioral features extracted in this study. More investigations are needed to answer whether they underperformed because the current state-of-the-art models cannot capture enough information in remote interviews, or behavioral signals are not as useful as physiological signals in telehealth settings, even for human experts.

Among behavioral modalities, overall facial and language patterns led to better classification performance than patterns derived from audio, although the latter resulted in a comparable performance in detecting self-rated depression. While overall

facial and language patterns led to similar levels of performance, it is worth noting that they performed very differently in different tasks, suggesting the same modality might perform differently for different mental health assessment tasks. For example, facial expression dynamics were much more useful in detecting self-rated anxiety than self-rated depression, yet similarly useful in detecting clinical MDD. On the contrary, Language embedding was more powerful in detecting self-rated disorders than clinically diagnosed disorders. These findings caution us on translating and interpreting results found using self-rated or self-reported scales directly for clinical applications, where the categorization criteria and process are different, in addition to the subject distribution shift, which was not shown in this study (as they were evaluated on the same group of subjects).

B. The Use of Foundation Models

Foundation models have gained enormous popularity in the last few years with the rapid development of pre-training and self(semi)-supervised training methods [94], [95], especially since the release of OpenAI's ChatGPT. LLMs, along with visual [53], audio [54], and multimodal [96] foundation models were widely applied in many disciplines, including the mental health domain, but primarily limited to language analyses and self-rated (self-reported) conditions [97], [98]. By comparing the direct use of unimodal foundation model-generated embedding to manually defined features from the same modalities, we showed how they perform in more clinically-relevant tasks under the telehealth settings.

The statistics of the visual embedding from DinoV2 were not at all useful in detecting mental health disorders. This finding was partially expected because the majority of extracted general visual representations would be more relevant to the texture and appearance of the face, especially after averaging, while the dynamics of the high-dimensional embedding would be hard to find with a limited number of recordings (discussed in more details in Section V-C below). Preliminary results using other vision foundation models in this dataset did not show better classification performances either, including using models tuned for facial representation ("FaRL" [99]) and for facial video representations ("MARLIN" [100]).

Although audio embedding from WavLM marginally outperformed acoustic features in our experiments, it demonstrates the potential of using the general audio embedding from a more diversely pre-trained audio foundation model in datasets with more subjects. Interestingly, general text embedding of the entire semi-structured interview from LLAMA-65B performed similarly when compared to sentiment analyses considering the extremely high feature dimensionality and the small number of recordings. With the rapid development of LLMs and the inclusion of more diverse training texts, such as the recent release of LLAMA2 [101], general LLMs could potentially outperform fine-tuned task-specific LLMs in mental health assessment tasks in the future.

C. Limitations and Future Directions

Several limitations of this study need to be acknowledged, as they provide valuable insights into the boundaries of our findings and potential directions for future studies.

First, the number of subjects ($n = 73$) and their heterogeneity might limit our findings' generalizability. While the number of subjects will grow as we keep collecting data following our defined protocol [13], the heterogeneity issue might not be easily addressed. Although we recruited subjects with clear inclusion/exclusion criteria and further excluded subjects after the interview if they did not fall into our criteria, the intrinsic nature of high comorbidity levels in different mental disorders makes it difficult to recruit a "clean" cohort of subjects with clear diagnoses of a single type of disorder. Another heterogeneity comes from medication status, which has been known to affect both behaviors and physiology of the patients [102], [103], [104]. Nevertheless, we believe the heterogeneity could be partially addressed as the number of subjects grows because analyses of smaller and more well-defined groups, for which we do not currently have enough samples, could be performed. As the number of subjects grows, models used for feature extractions could potentially be fine-tuned on the target population instead of being only trained on open-access datasets, which could further close the gap in identifying the most pertinent features from the target population.

Second, potential bias in the feature extraction and subject categorization processes might exist. The facial expression model used in this study was evaluated in our previous research [57], but the features from other modalities were extracted using open-access models that may bias towards certain demographic groups, leading to potential skew in the findings. For example, LLAMA is reported to be biased in religion, age, gender, and other aspects as it was trained with internet-crawled data [55]. A thorough bias analysis must be performed in a future study, and on a larger cohort before applying it clinically. Additionally, the subject categorization in this study may potentially contain inaccuracy or bias due to the limitation of the diagnostic process. While we partially addressed the issue by using both self-reported and clinician-rated measures and analyzing their relationships, future studies are needed for the direct investigation of this challenge. For example, the use of evaluations from multiple clinicians and a complete review of medical records may result in more accurate categorizations [105], [106], and specifically designed learning methods could be used to address the presence of noisy labels [107]. Such approaches would require reinterviewing each subject, which is costly and time-consuming, and would necessarily reduce the size of the cohort we have recruited.

Third, the unimodal and multimodal classification and fusion methods used in this study could be improved given a larger and more densely labeled dataset. Only one label (per task) was available for the entire recording, which made it difficult to apply temporal models such as recurrent neural networks or transformers to directly classify high-dimensional time series with thousands to tens of thousands of steps. Similarly, a multimodal transformer could be potentially used for fusion, provided the label sparsity challenge is addressed. A potential future direction is to label the entire recording more frequently in time. For example, simple measurements like self-rated or clinician-rated levels of distress could be adopted. Another potential direction is to utilize the potential improvement in pre-trained foundation models. For instance, LLMs with larger context windows might enable few-shot classification by including a few examples of transcripts and categorizations in the prompt.

Finally, the interpretability and explainability of the features and models remain unexplored [108]. They are key to fostering the trust of patients and clinicians and pushing the final clinical adoption. Interpretable machine learning methods [109] could

be applied to explain which modality, which feature, and which temporal section contribute to the system outputs. Additionally, visualization and reporting through dashboards and text summaries could be beneficial for clinicians and patients to understand better what was assessed and measured.

D. Potential Clinical Applications

With larger and more diverse samples, we see considerable potential clinical utility for the proposed multimodal objective assessment approach (and future applications informed by this technology) in several areas: 1) deepening understanding of psychopathology and outward manifestations of symptoms, 2) utility for diagnostic purposes, 3) assessing changes in symptoms longitudinally for the same patient, and 4) for patient self-report, engagement, and empowerment. First, this technology has the potential to better objectify and quantify core signs and symptoms of certain mental health conditions, like affect-flattening or tangential speech. Second, this technology has the potential to augment the initial diagnostic process for clinicians in both research and clinical settings. Developing real-time reporting of digital biomarker outputs in the form of a dashboard may help clinicians may hone into a certain line of clinical questions to better help establish a diagnosis. This technology may have a role in reducing bias and discrimination in the diagnostic process, as currently, the preponderance of evidence suggests that Black/African American individuals and Hispanic individuals are disproportionately diagnosed with psychotic disorders [110]. In time, combining digital biomarkers in addition to other blood-based and imaging markers, could play a potential role in subtyping mental health conditions according to treatment response or identifying individuals at risk who might develop the condition [111]. Third, applications of this technology can help clinicians and researchers to assess changes in symptoms over time for the same patient. This is crucial for the health care team to understand if the treatment plan is working and may help to accelerate measurement-based care efforts and overcome some of the barriers to implementation [112]. Additionally, accurate assessment is the cornerstone of clinical research studies, which ultimately determines whether new treatments are approved, and unreliable assessments can have significant consequences to the study and to the field more broadly [113]. Fifth, automated systems can provide quality assurance and feedback to clinicians on how well *they* performed during the interview, and potentially areas where their technique may be improved. Finally, future applications informed by this technology can play an important role in empowering patients to participate in self-assessment and ongoing monitoring of their symptoms. Such applications may help to improve the accessibility/timeliness of assessments and potentially reduce stigma around mental health [114].

References

[1] F. Charlson et al., "New WHO prevalence estimates of mental disorders in conflict settings: A systematic review and meta-analysis," *Lancet*, vol. 394, no. 10194, pp. 240–248, Jul. 2019.

[2] T. Vos et al., "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019," *Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[3] Substance Abuse and Mental Health Services Administration, "Projections of national expenditures for treatment of mental and substance use disorders, 2010–2020," U.S. Dept. Health Hum. Serv., Washington, DC, USA, HHS Publication Rep. SMA-14-4883 2014. [Online]. Available: https://store.samhsa.gov/product/Projections-of-National-Expenditures-for-Treatment-of-Mental-and-Substance-Use-Disorders-2010-2020/SMA14-4883

[4] Mental Health America, "Access to care data 2022," 2018. Accessed: Jul. 23, 2023. [Online]. Available: https://mhanational.org/issues/2022/mental-health-america-access-care-data

[5] V. N. Vahia, "Diagnostic and statistical manual of mental disorders 5: A quick glance," *Indian J. Psychiatry*, vol. 55, no. 3, 2013, Art. no. 220.

[6] World Health Organization, *International Statistical Classification of Diseases and Related Health Problems: Alphabetical Index*, vol. 3. Geneva, Switzerland: World Health Organization, 2004.

[7] D. E. Clarke et al., "DSM-5 field trials in the United States and Canada, Part I: Study design, sampling strategy, implementation, and analytic approaches," *Amer. J. Psychiatry*, vol. 170, no. 1, pp. 43–58, 2013.

[8] A. Aboraya, "Clinicians' opinions on the reliability of psychiatric diagnoses in clinical settings," *Psychiatry*, vol. 4, no. 11, 2007, Art. no. 31.

[9] H. N. Garb, "Race bias and gender bias in the diagnosis of psychological disorders," *Clin. Psychol. Rev.*, vol. 90, 2021, Art. no. 102087.

[10] R. L. Spitzer et al., "A brief measure for assessing generalized anxiety disorder: The GAD-7," *Arch. Intern. Med.*, vol. 166, no. 10, pp. 1092–1097, 2006.

[11] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001.

[12] E. I. Fried, J. K. Flake, and D. J. Robinaugh, "Revisiting the theoretical and methodological foundations of depression measurement," *Nature Rev. Psychol.*, vol. 1, no. 6, pp. 358–368, 2022.

[13] R. O. Cotes et al., "Multimodal assessment of schizophrenia and depression utilizing video, acoustic, locomotor, electroencephalographic, and heart rate technology: Protocol for an observational study," *JMIR Res. Protoc.*, vol. 11, no. 7, Jul. 2022, Art. no. e36417.

[14] Z. Jiang, S. Harati, A. Crowell, H. S. Mayberg, S. Nemati, and G. D. Clifford, "Classifying major depressive disorder and response to deep brain stimulation over time by analyzing facial expressions," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 2, pp. 664–672, Feb. 2021.

[15] G. Stratou et al., "Automatic nonverbal behavior indicators of depression and PTSD: The effect of gender," *J. Multimodal User Interfaces*, vol. 9, no. 1, pp. 17–29, 2015.

[16] E. G. Pintelas, T. Kotsilieris, I. E. Livieris, and P. Pintelas, "A review of machine learning prediction methods for anxiety disorders," in *Proc. 8th Int. Conf. Softw. Devlop. Technol. Enhancing Accessibility Fighting Info-Exclusion*, 2018, pp. 8–15.

[17] Z. Jiang et al., "Utilizing computer vision for facial behavior analysis in schizophrenia studies: A systematic review," *PLoS One*, vol. 17, no. 4, Apr. 2022, Art. no. e0266828.

[18] E. Reinertsen et al., "Continuous assessment of schizophrenia using heart rate and accelerometer data," *Physiol. Meas.*, vol. 38, no. 7, 2017, Art. no. 1456.

[19] S. Harati, A. Crowell, Y. Huang, H. Mayberg, and S. Nemati, "Classifying depression severity in recovery from major depressive disorder via dynamic facial features," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 3, pp. 815–824, Mar. 2020.

[20] Z. Jiang et al., "Disentangling visual exploration differences in cognitive impairment," *IEEE Trans. Biomed. Eng.*, early access, Nov. 9, 2023, doi: 10.1109/TBME.2023.3330976.

[21] S. Harati, A. Crowell, H. Mayberg, and S. Nemati, "Depression severity classification from speech emotion," in *Proc. IEEE 40th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2018, pp. 5763–5766.

[22] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, Jul. 2000.

[23] A. Qayyum, I. Razzak, M. Tanveer, M. Mazher, and B. Alhaqbani, "High-density electroencephalography and speech signal based deep framework for clinical depression diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 4, pp. 2587–2597, Jul./Aug. 2023.

[24] U. Ahmed, J. C.-W. Lin, and G. Srivastava, "Graph attention-based curriculum learning for mental healthcare classification," *IEEE J. Biomed. Health Inform.*, early access, May 8, 2023, doi: 10.1109/JBHI.2023.3274486.

[25] A. S. Cakmak et al., "Classification and prediction of post-trauma outcomes related to PTSD using circadian rhythm changes measured via wrist-worn research watch in a large longitudinal cohort," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 8, pp. 2866–2876, Aug. 2021.

[26] N. Palmius et al., "Detecting bipolar depression from geographic location data," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1761–1771, Aug. 2017.

[27] G. Valenza et al., "Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 5, pp. 1625–1635, Sep. 2014.

[28] J. A. Boscarino and J. Chang, "Electrocardiogram abnormalities among men with stress-related psychiatric disorders: Implications for coronary heart disease and clinical research," *Ann. Behav. Med.*, vol. 21, no. 3, pp. 227–234, 1999.

[29] U. R. Acharya et al., "Computer-aided diagnosis of depression using EEG signals," *Eur. Neurol.*, vol. 73, no. 5/6, pp. 329–336, 2015.

[30] Y. Zhang et al., "Identification of psychiatric disorder subtypes from functional connectivity patterns in resting-state electroencephalography," *Nature Biomed. Eng.*, vol. 5, no. 4, pp. 309–323, 2021.

[31] J. H. Yoon et al., "Automated classification of fMRI during cognitive control identifies more severely disorganized subjects with schizophrenia," *Schizophrenia Res.*, vol. 135, no. 1–3, pp. 28–33, 2012.

[32] Y. Du et al., "NeuroMark: An automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders," *NeuroImage: Clin.*, vol. 28, 2020, Art. no. 102375.

[33] H. Song et al., "Automatic schizophrenic discrimination on fNIRS by using complex brain network analysis and SVM," *BMC Med. Inform. Decis. Mak.*, vol. 17, pp. 1–9, 2017.

[34] I. Moura et al., "Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: A systematic literature review," *J. Biomed. Inform.*, vol. 138, 2022, Art. no. 104278.

[35] E. Garcia-Ceja et al., "Mental health monitoring with multimodal sensing and machine learning: A survey," *Pervasive Mobile Comput.*, vol. 51, pp. 1–26, 2018.

[36] R. Gupta et al., "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proc. ACM Int. Workshop Audio/Vis. Emotion Challenge*, 2014, pp. 33–40.

[37] S. Ghosh, M. Chatterjee, and L.-P. Morency, "A multimodal context-based approach for distress assessment," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2014, pp. 240–246.

[38] X. Zhang, J. Shen, Z. u. Din, J. Liu, G. Wang, and B. Hu, "Multimodal depression detection: Fusion of electroencephalography and paralinguistic behaviors using a novel strategy for classifier ensemble," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 6, pp. 2265–2275, Nov. 2019.

[39] D. M. Mann et al., "COVID-19 transforms health care through telemedicine: Evidence from the field," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 7, pp. 1132–1135, 2020.

[40] J. J. Moffatt and D. S. Eley, "The reported benefits of telehealth for rural Australians," *Australian Health Rev.*, vol. 34, no. 3, pp. 276–281, 2010.

[41] N. R. Cunningham et al., "Addressing pediatric mental health using telehealth during coronavirus disease-2019 and beyond: A narrative review," *Academic Pediatrics*, vol. 21, no. 7, pp. 1108–1117, 2021.

[42] S. Sultana and J. A. Pagán, "Use of telehealth to address depression and anxiety in low-income us populations: A narrative review," *J. Primary Care Community Health*, vol. 14, 2023, Art. no. 21501319231168036.

[43] J. Wright-Berryman et al., "Virtually screening adults for depression, anxiety, and suicide risk using machine learning and language from an open-ended interview," *Front. Psychiatry*, vol. 14, 2023, Art. no. 1143175.

[44] A. Abbas et al., "Computer vision-based assessment of motor functioning in schizophrenia: Use of smartphones for remote measurement of schizophrenia symptomatology," *Digit. Biomarkers*, vol. 5, no. 1, pp. 29–36, 2021.

[45] F. Matcham et al., "Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): Recruitment, retention, and data availability in a longitudinal remote measurement study," *BMC Psychiatry*, vol. 22, no. 1, 2022, Art. no. 136.

[46] D. V. Sheehan et al., "The mini-international neuropsychiatric interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10," *J. Clin. Psychiatry*, vol. 59, no. 20, pp. 22–33, 1998.

[47] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[48] R. J. Moretti and E. D. Rossini, *Comprehensive Handbook of Psychological Assessment, Vol. 2: Personality Assessment*. Hoboken, NJ, USA: Wiley, 2004.

[49] A. L. Benton, K. deS, and A. B. Sivan, *Multilingual Aphasia Examination*. Bangalore, India: AJA Associates, 1994.

[50] H. Goodglass and E. Kaplan, *The Assessment of Aphasia and Related Disorders*. Philadelphia, PA, USA: Lea & Febiger, 1972.

[51] S. R. Cohen et al., "Measuring the quality of life of people at the end of life: The McGill quality of life questionnaire—revised," *Palliat. Med.*, vol. 31, no. 2, pp. 120–129, 2017.

[52] G. Boccignone et al., "pyVHR: A python framework for remote photoplethysmography," *PeerJ Comput. Sci.*, vol. 8, 2022, Art. no. e929.

[53] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[54] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[55] H. Touvron et al., "LLAMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[56] Y. Liu et al., "RoBERTa: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.

[57] Z. Jiang et al., "Automated analysis of facial emotions in subjects with cognitive impairment," *PLoS One*, vol. 17, no. 1, Jan. 2022, Art. no. e0262527.

[58] J. Deng et al., "RetinaFace: Single-stage dense face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5203–5212.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[60] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5525–5533.

[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[62] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.

[63] O. Langner et al., "Presentation and validation of the radboud faces database," *Cogn. Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[64] Z. Shao et al., "JAA-Net: Joint facial action unit detection and face alignment via adaptive attention," *Int. J. Comput. Vis.*, vol. 129, pp. 321–340, 2021.

[65] P. Ekman and W. V. Friesen, "Facial action coding system," *Environ. Psychol. Nonverbal Behav.*, 1978, doi: 10.1037/t27734-000.

[66] X. Zhang et al., "BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, 2014.

[67] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[68] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[69] S. Seyedi et al., "Using HIPAA (health insurance portability and accountability act)–compliant transcription services for virtual psychiatric interviews: Pilot comparison study," *JMIR Ment. Health*, vol. 10, Oct. 2023, Art. no. e48517.

[70] L. Corbin et al., "A comparison of linguistic patterns between individuals with current major depressive disorder, past major depressive disorder, and those without major depressive disorder in a virtual, psychiatric research interview," *J. Affect. Disord. Rep.*, vol. 14, 2023, Art. no. 100645.

[71] J. Hartmann, "Emotion English DistilRoBERTa-base," 2022. [Online]. Available: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/

[72] J. Hartmann et al., "More than a feeling: Accuracy and application of sentiment analysis," *Int. J. Res. Marketing*, vol. 40, no. 1, pp. 75–87, 2023.

[73] F. Ahmad et al., "A deep learning architecture for psychometric natural language processing," *ACM Trans. Inf. Syst.*, vol. 38, no. 1, pp. 1–29, 2020.

[74] T. Giannakopoulos, "pyAudioAnalysis: An open-source Python library for audio signal analysis," *PLoS One*, vol. 10, no. 12, 2015, Art. no. e0144610.

[75] S.-W. Yang et al., "SUPERB: Speech processing universal performance benchmark," *22nd Ann. Conf. Int. Speech Commun. Assoc.*, pp. 3161–3165, 2021.
[76] G. Boccignone, D. Conte, V. Cuculo, A. D'Amelio, G. Grossi, and R. Lanzarotti, "An open framework for remote-PPG methods and their assessment," *IEEE Access*, vol. 8, pp. 216083–216103, 2020.
[77] C. A. Casado and M. B. López, "Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 11, pp. 5530–5541, Nov. 2023.
[78] A. N. Vest et al., "An open source benchmarked toolbox for cardiovascular waveform and interval analysis," *Physiol. Meas.*, vol. 39, no. 10, 2018, Art. no. 105004.
[79] S. Linderman, B. Antin, D. Zoltowski, and J. Glaser, "SSM: Bayesian learning and inference for state space models," Oct. 2020. [Online]. Available: https://github.com/lindermanlab/ssm
[80] H. Ma, M. Cai, and H. Wang, "Emotional blunting in patients with major depressive disorder: A brief non-systematic review of current research," *Front. Psychiatry*, vol. 12, 2021, Art. no. 792960.
[81] F. Trémeau et al., "Facial expressiveness in patients with schizophrenia compared to depressed patients and nonpatient comparison subjects," *Amer. J. Psychiatry*, vol. 162, no. 1, pp. 92–101, 2005.
[82] L. M. Bylsma, B. H. Morris, and J. Rottenberg, "A meta-analysis of emotional reactivity in major depressive disorder," *Clin. Psychol. Rev.*, vol. 28, no. 4, pp. 676–691, 2008.
[83] H. Davies et al., "Facial expression to emotional stimuli in non-psychotic disorders: A systematic review and meta-analysis," *Neurosci. Biobehavioral Rev.*, vol. 64, pp. 252–271, 2016.
[84] G. Valenza et al., "Mood states modulate complexity in heartbeat dynamics: A multiscale entropy analysis," *Europhysics Lett.*, vol. 107, no. 1, 2014, Art. no. 18003.
[85] L. Zhao et al., "Cardiorespiratory coupling analysis based on entropy and cross-entropy in distinguishing different depression stages," *Front. Physiol.*, vol. 10, 2019, Art. no. 359.
[86] K. Schultebraucks et al., "Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood," *Psychol. Med.*, vol. 52, no. 5, pp. 957–967, 2022.
[87] Y. Xing et al., "Task-state heart rate variability parameter-based depression detection model and effect of therapy on the parameters," *IEEE Access*, vol. 7, pp. 105701–105709, 2019.
[88] S. Byun et al., "Detection of major depressive disorder from linear and nonlinear heart rate variability features during mental task protocol," *Comput. Biol. Med.*, vol. 112, 2019, Art. no. 103381.
[89] K. M. Hasib, M. R. Islam, S. Sakib, M. A. Akbar, I. Razzak, and M. S. Alam, "Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey," *IEEE Trans. Comput. Social Syst.*, vol. 10, no. 4, pp. 1568–1586, Aug. 2023.
[90] K. Yang, R. Y. Lau, and A. Abbasi, "Getting personal: A deep learning artifact for text-based measurement of personality," *Inf. Syst. Res.*, vol. 34, no. 1, pp. 194–222, 2023.
[91] J. W. Weeks et al., ""The sound of fear": Assessing vocal fundamental frequency as a physiological indicator of social anxiety disorder," *J. Anxiety Disord.*, vol. 26, no. 8, pp. 811–822, 2012.
[92] A. J. Grippo and A. K. Johnson, "Stress, depression and cardiovascular dysregulation: A review of neurobiological mechanisms and the integration of research from preclinical disease models," *Stress*, vol. 12, no. 1, pp. 1–21, 2009.
[93] X. Zang et al., "End-to-end depression recognition based on a one-dimensional convolution neural network model using two-lead ECG signal," *J. Med. Biol. Eng.*, vol. 42, no. 2, pp. 225–233, 2022.
[94] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer.*, vol. 1, p. 2, 2019.
[95] T. Brown et al., "Language models are few-shot learners," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
[96] OpenAI, "GPT-4 technical report," 2023, *arXiv:2303.08774*.
[97] C. Lau, X. Zhu, and W.-Y. Chan, "Automatic depression severity assessment with deep learning using parameter-efficient tuning," *Front. Psychiatry*, vol. 14, 2023, Art. no. 1160291.
[98] N. Farruque et al., "Depression symptoms modelling from social media text: A semi-supervised learning approach," 2022, *arXiv:2209.02765*.
[99] Y. Zheng et al., "General facial representation learning in a visual-linguistic manner," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18697–18709.
[100] Z. Cai et al., "MARLIN: Masked autoencoder for facial video representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1493–1504.
[101] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.
[102] W. K. Goodman, T. K. Murphy, and E. A. Storch, "Risk of adverse behavioral effects with pediatric use of antidepressants," *Psychopharmacology*, vol. 191, pp. 87–96, 2007.
[103] J. S. Buyukdura, S. M. McClintock, and P. E. Croarkin, "Psychomotor retardation in depression: Biological underpinnings, measurement, and treatment," *Prog. Neuro-Psychopharmacology Biol. Psychiatry*, vol. 35, no. 2, pp. 395–409, 2011.
[104] J. P. Halper and J. J. Mann, "Cardiovascular effects of antidepressant medications," *Brit. J. Psychiatry*, vol. 153, no. S3, pp. 87–98, 1988.
[105] S. Nasiri et al., "Exploiting labels from multiple experts in automated sleep scoring," *Sleep*, vol. 46, no. 5, 2023, Art. no. zsad034.
[106] M. Nasir, B. Baucom, P. Georgiou, and S. Narayanan, "Redundancy analysis of behavioral coding for couples therapy and improved estimation of behavior from noisy annotations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 1886–1890.
[107] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8135–8153, Nov. 2023.
[108] D. S. Johnson, O. Hakobyan, and H. Drimalla, "Towards interpretability in audio and visual affective machine learning: A review," 2023, *arXiv:2306.08933*.
[109] G. Stiglic et al., "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 5, 2020, Art. no. e1379.
[110] R. C. Schwartz and D. M. Blankenship, "Racial disparities in psychotic disorder diagnosis: A review of empirical literature," *World J. Psychiatry*, vol. 4, no. 4, 2014, Art. no. 133.
[111] D. R. Goldsmith et al., "An update on promising biomarkers in schizophrenia," *Focus*, vol. 16, no. 2, pp. 153–163, 2018.
[112] C. C. Lewis et al., "Implementing measurement-based care in behavioral health: A review," *JAMA Pychiatry*, vol. 76, no. 3, pp. 324–335, 2019.
[113] S. Berendsen et al., "Burying our heads in the sand: The neglected importance of reporting inter-rater reliability in antipsychotic medication trials," *Schizophrenia Bull.*, vol. 46, no. 5, pp. 1027–1029, 2020.
[114] M. E. Rodríguez-Rivas et al., "Innovative technology–based interventions to reduce stigma toward people with mental illness: Systematic review and meta-analysis," *JMIR Serious Games*, vol. 10, no. 2, 2022, Art. no. e35099.