Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders

Bochao Zou, *Member, IEEE*, Jiali Han, Yingxue Wang, Rui Liu, Shenghui Zhao, Lei Feng, Xiangwen Lyu, and Huimin Ma*

Abstract—Depression is a common psychiatric disorder worldwide. However, in China, a considerable number of patients with depression are not diagnosed, and most of them are not aware of their depression. Despite increasing efforts, the goal of automatic depression screening from behavioral indicators has not been achieved. A major limitation is the lack of available multimodal depression corpus in Chinese since linguistic knowledge is crucial in clinical practice. Therefore, we first carried out a comprehensive survey with psychiatrists from a renowned psychiatric hospital to identify key interview topics which are highly related to the diagnosis of depression. Then, a semi-structural interview study was conducted over a year with subjects who have undergone clinical diagnosis and professional assessment. After that, Visual, acoustic, and textual features were extracted and analyzed between the two groups, statistically significant differences were observed in all three modalities. Benchmark evaluations of both single modal and multimodal fusion methods of depression assessment were also performed. A multimodal transformer-based fusion approach achieved the best performance. Finally, the proposed Chinese Multimodal Depression Corpus (CMDC) was made publicly available after de-identification and annotation. Hopefully, the release of this corpus would promote the research progress and practical applications of automatic depression screening.

Index Terms—Affective computing; depressive disorder; multimodal corpus; semi-structural interview

1 Introduction

Depression, otherwise known as major depressive disorder (MDD), is a common psychiatric disorder that negatively impacts a person's way of thinking, feeling, and behavior [1]. With the rapid development of society and the increasing pressure on people's work and life, depression has become one of the most common and serious mental diseases worldwide [2]. Up to now, the number of patients with depression in China has increased to 95 million, becoming the country with the largest number of depressive patients in the world [3]. According to an epidemiological survey: In China, the lifetime prevalence rate of depression is about 6.9%, of which less than 10% of patients with depression have received professional assistance and treatment, and a considerable number of patients are not aware of their depression [4]. On the other hand, among the few patients who seek treatment in time, the first hospital most of them visit is not psychiatric hospitals nor the psychiatric department of general hospitals, which is easy to cause misdiagnosis, and ultimately delay the treatment. Consequently, the National Health Commission of China issued the first action plan for the prevention and control of depression, entitled "Action Plan for Explorations of Specialized Services for the Prevention and Treatment of Depression", on September 11, 2020, which includes the routine screening of depression throughout the country [5].

Screening and prevention of depression are of great significance. However, traditional questionnaire-based screening of depression is facing problems of lacking well-trained healthcare personnel since clinical interviewed-based screening is labor-intensive and self-evaluation questions lack accuracy [6]. Many symptoms of depression are considered observable [7]–[9]. The Diagnostic and Statistical Manual of Mental Disorders (DSM) is the standard of psychiatric diagnosis, which describes a series of audio-visual behavioral indicators of depression [10]. However, these indicators are often not fully considered when screening, diagnosing, and evaluating depression [11]. The assessment of depression relies almost entirely on patients' orally reported symptoms described in particular ques-

- *B. Zou and H. Ma are with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, 100083, China. E-mail: zoubochao@ustb.edu.cn, mhmpub@ustb.edu.cn (corresponding author).*
- *J. Han, R. Liu, and L. Feng are with the National Clinical Research Centre of Mental Disorders, Beijing Anding Hospital of Capital Medical University, Beijing, 100088, China. E-mail: jlhan@mail.ccmu.edu.cn, ruiliu@ccmu.edu.cn, flxlm@ccmu.edu.cn.*
- *Y. Wang and X. Lyu are with the National Engineering Laboratory for Risk Perception and Prevention, Beijing, 100041, China. E-mail: wangyingxue@cetc.com.cn, lvxiangwen@cetc.com.cn.*
- *S. Zhao is with the School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, China. E-mail: shzhao@bit.edu.cn.*

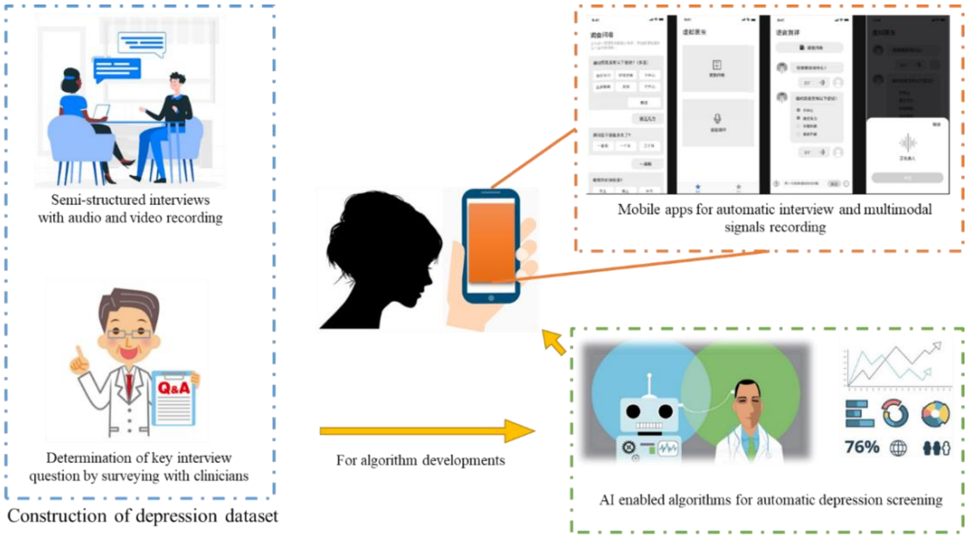

Semi-structured interviews
with audio and video recording
Determination of key interview
question by surveying with clinicians
Construction of depression dataset
For algorithm developments
Mobile apps for automatic interview and multimodal
signals recording
AI enabled algorithms for automatic depression screening
Q&A
76%

Fig. 1. Blueprint for automatic pre-screening of depressive disorders through mobile application with build-in AI algorithms.

tionnaires [12], such as the clinician-administered Hamilton Rating Scale for Depression (HAMD) [13] and the self-report Patient Health Questionnaire (PHQ-9) [14]. Although these tools are useful, they neither include visual, acoustic, or textual indicators of depression. To overcome this limitation, recent advancements in techniques for automatic analysis of human behaviors, such as computer vision, speech signal processing, natural language understanding, and multimodal learning could play an important role [15], [16].

There are considerable research interests in developing tools to analyze the video [17], audio [18]–[20], and text [21] content of clinical interviews automatically as a means of medical aided diagnosis [19], [22]–[25]. Despite increasing efforts, the goal of automatically, reliably, and objectively screening of depression from behavioral indicators has not been achieved [26]. Because of the huge population and the prevalence rate of depressive disorder in China, the construction of a multimodal depression corpus with semi-structural interviews in Chinese would be helpful to promote the auxiliary screening of depression based on information technologies, which is expected to realize the automatic primary screening of depression and reduce the medical burden of society.

One challenge of automatic depression screening is the lack of available multimodal datasets which contain behavior observations of patients with clinically validated depression [27], [28]. Therefore, in this paper, we proposed the Chinese multimodal depression corpus (CMDC), which is a publicly available multimodal Chinese depression dataset based on clinically validated semi-structural interviews for AI-enabled depression screening, diagnosis, and assessment. The contributions of this paper are as follows:

- Key interview topics for the development of automatic depression screening tools are identified. We performed a comprehensive survey of MDD majored clinicians from a renowned psychiatric hospital in China to identify key interview questions which are highly related to the diagnosis of depression and can be used for AI-enabled automatic depression screening, as shown in Fig. 1.
- We conducted semi-structural interviews based on key interview questions with participants of MDD and Health Control (HC) subjects who have undergone clinical diagnosis and professional assessment of symptom severity. During the interviews, the video and audio of participants were recorded simultaneously, and the audio was transcribed into text through automatic transcription tools and proofreading.
- Significant features of visual, acoustic, and textual modalities between MDD and HC groups are revealed with statistical analysis. These significant differences among features confirm the feasibility of automatic depression analysis with machine learning methods.
- Comprehensive benchmark evaluations are conducted on the proposed dataset to provide a basis for any follow-up depression assessment research. A multimodal transformer-based fusion approach is applied in the domain of depression analysis which achieved the best result during evaluation.
- The proposed Chinese multimodal dataset was made publicly available after annotation and de-identification[1]. To distribute the dataset publicly, we extract the features of key questions related videos and audios with open-source toolkits to eliminate personal information.

We believe that the release of this dataset will promote the research progress and practical applications of depression screening and assessment with core affective computing technologies, which could have significant scientific value and broad application for societal questions of mental health.

This paper is organized as follows. Section 2 reviews re-

[1] Data and code are available at: https://github.com/CMDC-corpus/CMDC-Baseline

lated work on depression-related behavioral patterns, assessment methods, and multimodal datasets. Section 3 and 4 describe the experiment details and feature extraction procedures, respectively. Section 5 presents the results of statistical analysis and benchmark evaluations. Finally, Section 6 concludes the paper.

2 Related Work

2.1 Depression related Behavioral Patterns

Previous studies have shown significant differences between MDD and HC subjects in several aspects of behavioral signals, such as visual, acoustic, and textual cues observed during interviews [29], [30]. Depression could be depicted in patients' appearance (facial expression and body posture) [31]. Both global and local facial features, such as eyes and mouth area, are of particular interest for depression assessment. For instance, the eye movements of depressed patients were shown to have statistically significant differences from the HC group [32]. It was reported that larger downward angles of gaze, shorter average duration, and less intensity of smiles are the most significant facial cues of depression [33]. Findings regarding psychomotor disturbance of bipolar disorders showed an increase in reaction time in saccadic tasks [34]. Acoustic features were also found consistently different between MDD and HC with large effect sizes [35], [36]. Decreased speech rate and longer reaction time were found in depressed subjects [37]. Pitch and loudness were widely used features in depression detection studies and have been shown to have a negative relationship with depression severity [38]–[40]. There are studies showing that textual features also play an important role in depression detection, indicating the importance of semantic information [41]–[43]. These explorations facilitate the interpretation of depression behavior since it is obtained through multimodal behavior analysis of clinically matched control depression dataset. Thus, it is important to identify the most meaningful patterns of depressive behavior since behavior is associated with depression-related symptoms in psychiatry studies [27].

2.2 Depression Assessment with Behavioral Indicators

There is an increasing number of studies on behavior indicators-based depression detection, and the research trend extends from traditional manually designed features to more advanced deep learning methods [19], [34], [43]–[47]. For visual modality, AUs, eye gaze, head poses, and facial landmarks are common features for depression-related analysis and can be combined with machine/deep learning tools for depression detection. Pampouchidou et al. [34] gave a systematic review of visual cues based methods. Recent studies began to pay more attention to the dynamics of visual information with spatial-temporal modeling architectures [26] and graph neural networks [48]. As to audio modality, Mel filters, spectrograms, and emotion feature sets, e.g. eGemaps, were widely adopted [49], [50]. Audio models pretrained on large scale datasets were deployed as features extractors to cope with the bottleneck of small sample number, such as in [19], they used pretrained VGGish [51] to extract the sentence-level vector of each speech segment, and then LSTM network with self-attention mechanism was integrated to train the downstream classification task. A latest review of deep learning based depression analysis with audiovisual cues can be found in [52]. Semantic information is also of great importance in depression detection, previous studies have suggested superior performance of textual features [53]. Contextual sentence embeddings, such as BERT [54], were shown to achieve better performance than that of word-level [41]. However, a recent study which used graph neural network to form the embedding of specific nodes of word vectors showed that their method outperforms previous state-of-art methods by a substantial margin [46].

The results of several years' Audio/Visual Emotion Challenge and Workshop (AVEC) showed that methods based on multimodal fusion usually achieved better results [49], [55], [56]. In terms of fusion paradigm, early fusion, feature-level fusion, and late fusion were all explored by relevant research [47]. With the success of transformers in natural language understanding and computer vision tasks, transformer-based fusion methods have also been proposed as fusion methods and demonstrated their advantages with temporal data by automatically aligning and capturing complementary features [57]. In summary, a wide range of studies have been conducted in the field of depression detection based on multimodal features, and considerable progress has been made in performance, which provides a guarantee for the practicability of preliminary screening for depression based on semi-structured interviews.

2.3 Multimodal Datasets for Depression Assessment

Well-labeled multimodal recordings of clinically relevant behavioral differences between depressive and healthy subjects are necessary for an automatic screening system to train classifiers [26]. Clinical datasets are hard to construct since the difficulty in participant recruiting, and are usually public unavailable due to the confidentiality of patient data. Table I shows a summary of interview-based datasets for depression assessment. We put these together to conduct a thorough analysis to highlight the strength of the proposed dataset.

Among them, the Distress Analysis Interview Corpus (DAIC-WOZ) [58], University of Pittsburgh depression dataset (Pitt) [7], and Black Dog Institute depression dataset [28] are the three influential depression datasets. Specifically, the BlackDog dataset was collected in a depression-specialized clinical research facility. Their interviews were conducted with specific open-ended questions, such as portrayal of occasions in their life that had excited critical feelings. Until now, the dataset has not been made public yet. The Pitt dataset contained 49 participants in a clinical trial for the treatment of depression [7]. All their participants met DSM-IV criteria for MDD. The severity of depression was evaluated with HAMD-17. This dataset is distributed upon request. The DAIC-WOZ dataset contains audio and facial features of depressed patients and control subjects. The expert evaluated HAMD-17 and self-report PHQ-8 scores are provided for each patient. This archive was also created from semi-structural interviews

TABLE 1
SUMMARY OF INTERVIEW-BASED DATASETS FOR DEPRESSION ASSESSMENT

Dataset	Participant source	Subject No. (MDD/HC)	Selection Criteria	Clinical Diagnosis	Interview Questions	Labels	Modality	Availability to Third Parties	Language
DAIC-WOZ [58]	depression, PTSD, and anxiety	189(-/-)	-	No	Open ended question	PHQ-8	A, V, T	Yes	English
Black Dog Institute [28]	Melancholia or MDD	60(30/30)	Clinical assessment	Yes (DSM)	Open ended question	QIDS-SR	A, V	No	English
Pittsburgh [7]	MDD	49(49/-)	HAMD>15	Yes (DSM)	From HAMD	HAMD	A, V	Upon request	English
Lin et al. [60]	depression and anxiety	35(18/17 for high & low distress)	PHQ-8⩾ 6.63	No	Peer-support interview questions	PHQ-8 and GAD-7	V, A	Upon request	English
Jiang et al. [61]	Depression	12(12/-)	a 50% decrease in HAMD score	Yes	Unstructured	HAMD	V	No	English
Guo et al. [62]	Depression	208(104/104)	PHQ-9⩾5	Yes	Based on depression scales	PHQ-9	A, V	No	Chinese
MODMA [63]	MDD	52(23/29)	PHQ-9⩾5	Yes (DSM)	Based on depression scales	PHQ-9	A, EEG	Yes	Chinese
Shen et al. [64]	Depression	162(30/132)	SDS⩾53	No	Randomly selected questions	SDS	A, T	Yes	Chinese
Proposed	**MDD**	**78(26/52)**	**HAMD-17 >17 or PHQ-9⩾9**	**Yes (DSM)**	**Semi structural questions identified with expert survey**	**HAMD & PHQ-9**	**A, V, T**	**Yes (De-identified data and features)**	**Chinese**

where research assistants or a computer agent asked a series of questions designed to identify depressive symptoms [58]. There are also depression detection datasets in the AVEC [59], which is a series of competitions that have been held for several years. In AVEC 2013, 2014, 2016, 2017, and 2019, there were sub-challenges in depression detection. The datasets in AVEC 2013 and 2014 were task-driven behavior observations of depressive people, which were not interview-based. For 2016, 2017, and 2019, they all employed the subset of DAIC-WOZ dataset.

Besides the above three datasets, Lin et al. [60] introduced a new audio-visual dataset containing full body videos for distress detection. Currently, only a few studies attempted to include the body modality which is worth exploring. In terms of dataset construction, their participants were recruited online without clinical diagnosis. Both depression and anxiety participants were included. The comorbidity of depression and anxiety makes it a big challenge to distinguish between these two. For our purpose, to develop prescreening of MDD, including patients with other psychotic disorders may introduce confounding factors. Visual and audio were recorded in their dataset but without text. Jiang et al.'s study [61] has a different research question with a cohort of 12 depressed patients aimed at assessing the recovery, as well as the response to deep brain stimulation treatment [65]. Their interview was unstructured and data are not available. There are also more early studies that are not included in Table 1 since they are not available anymore, such as ORYGEN [66] and MHMC [67], and they are not multimodal datasets, only containing video or audio.

In recent years, Chinese depression datasets were also developed by various studies. Guo et al. [62] proposed a large scale dataset (208 subjects) with audio and video recording while interviewing three categories of questions based on emotion polarity. But the authors claimed that the data would not be disclosed due to privacy issues. The MODMA [63] is a clinically validated and publicly available dataset. However, it only contains audio and EEG signals, while EEG is not feasible for preliminary screening. The newly published dataset by Shen et al. [64] recruited student volunteers from only one university which lacked a diversity of demographic characteristics. Their ground truth was from the Self-rating Depression Scale (SDS) [68]. A study with patients in China showed that SDS is less sensitive than PHQ-9 with statistically significant differences [69]. Moreover, the visual modality was not available in their dataset.

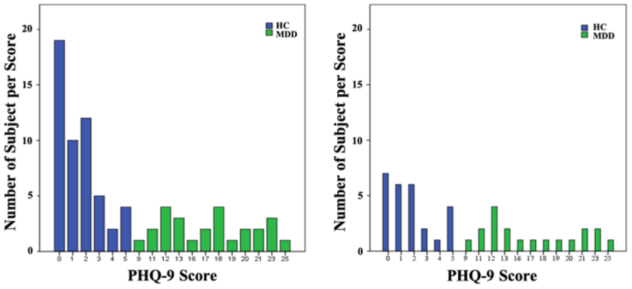
Considering the goal of preliminary screening of MDD in China, a clinically validated Chinese multimodal depression assessment dataset is certainly beneficial. As to inclusion/exclusion criteria, DAIC-WOZ has a range of depressive symptoms (depression, post-traumatic stress disorder, PTSD, and anxiety). Whether their participants met DSM was not considered, which we believe to be crucial since different psychotic disorders may show different behavior patterns. For the BlackDog dataset, they treated Melancholia and MDD patients as one class because of the relatively small sample size. Datasets of [61] and [62] do not state the diagnostic criteria they used. Datasets of [60] and [64] even do not have clinical diagnoses. Diagnostic criteria matter because depression is confusable with many non-depressive disorders. By using diagnostic criteria and paying attention to behavioral changes of depression, we can exclude other confounding factors. Besides, most of interview questions in previous studies were based on different questionnaires with various question numbers and were somehow random. The determination of the interview topic is key for prescreening which can ensure the accuracy and time consuming of the tool. For example, the number of questions would determine the time of screening and the robustness of the recognition algorithm will be beneficial from the structured interview.

Therefore, we conduct a comparably large-scale Chinese multimodal depression study with semi-structural interviews under clinically validated diagnosis. As shown in Table I, the strength of the proposed dataset are highlighted in the following aspects: first, rigorous inclusion/exclusion criterion: clinical diagnosed pure MDD patients as subjects; second, well-defined semi-structural interview questions through an extensive survey of MDD majored clinicians; third, publicly available multimodal (video, audio, and text) depression dataset in Chinese.

3 Experiment

3.1 Participants

This was a cross-sectional study conducted from Nov. 2018 to Jan. 2020. The MDD patients were recruited from Beijing Anding Hospital and the HCs were recruited from Beijing Institute of Technology. In total, Seventy-eight subjects participated this experiment, which included 26 MDD patients (8 males, 18 females) with a mean age of 24.1 (SD = 5.04, age range 19-30 yr), and 52 HC (17 males, 35 females) with a mean age of 30.5 (SD = 12.06, age range 20-60 yr). The research was approved by the Independent Medical of Ethics Committee Board of Beijing Anding Hospital (2019 No. 53). Written informed consent was obtained from each subject. Subjects were asked before the experiment if they would like to be video recorded, and the recorded video may be published in research papers without any personal information in the consent form. Among all subjects, as shown in Table 2, 45 subjects have consented to video recording (19 MDD (14 females, mean age=23.6, SD=3.45) and 26 HC (20 females, mean age=30.5, SD=11.92)), the rest were audio-recorded only. The distribution of PHQ-9 scores of 78 subjects, as well as 45 subjects with video recording, are shown in Fig. 2. The Mini International Neuropsychiatric Interview (MINI) was employed to obtain the diagnosis [70]. All MDD participants met DSM criteria for major depression as assessed by professional psychiatrists. Both HAMD-17 and PHQ-9 were assessed to serve as ground-truth labels for the development of automatic AI tools. PHQ-9 is a self-assessment questionnaire used to score nine DSM-IV criteria for depression. It is widely used as a tool for self-screening of depression. HAMD is a clinician-rated tool composed of 17 items to quantify the severity of depression. The HAMD assessed the severity by exploring mood, guilt feeling, suicidal ideation, insomnia, anxiety, and somatic symptoms. Interviewers were experts in HAMD. One thing to note in Table I is that different studies may have different criteria for depression and control grouping. For instance, studies of [62] and [63] adopted PHQ-9⩾5 for the depressed subject while Lin et al. [60] used PHQ-9⩾6.63. Lin et al. grouped their participants according to the public norm. A score between 5-9 suggests mild depression which may require only watchful waiting. Scores between 10-14 suggest moderate depression severity that patients should have a treatment plan [14]. Instead of HAMD > 15 for moderate to severe depression used in Pittsburgh [7] dataset. We deploy the following criterion recommended by [71]: A HAMD score of 18 or higher is generally considered to be moderate to severe depression; a score of 8-17 indicates mild depression; a score of 7 or less indicates remission. These cutoff scores are established with a large study of psychiatric outpatients with major depressive disorder and are currently adopted in the clinical scenario in China. A PHQ-9 of 9 or higher is semantically comparable to a HAMD score of 18 or higher. As suggested by clinical psychiatrists, these scores are adopted as the selection criterion in our study.



(a) (b)

Fig. 2. Distribution of PHQ-9 scores, (a) all 78 subjects, (b) 45 subjects with video recording

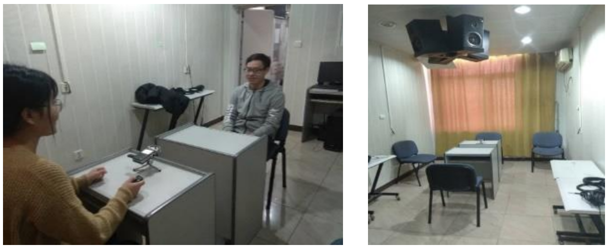


Fig. 3. Semi-structural interview setup.

TABLE 2
Summary of Participants

	MDD	HC
	All subjects	
No.	26	52
Age	24.1±5.04	30.5±12.1
Gender	8 M & 18 F	17 M & 35 F
	With video recording	
No.	19	26
Age	23.6±3.45	30.5±11.9
Gender	5 M & 14 F	6 M & 20 F

Only native Chinese-speaking participants were recruited to reduce the differences caused by different language acquisition. Due to the limited availability of MDD subjects, the HC group has more number of subjects and a broader range of ages. One can select a subset of the dataset if matched control were needed where the effect of gender

TABLE 3
KEY INTERVIEW TOPICS

No.	Questions
1	How are your appetite and weight change in the last two weeks?
2	How's your health recently?
3	Sleep-related questions*
4	How often do you communicate with your friends recently? What is your best friend's evaluation of you?
5	How about your memory recently? Do you often forget things?
6	How are you interested in your current study or job? How is your concentration in your daily life?
7	When do you feel tired or lack of energy? Does this happen frequently?
8	Have you ever thought of committing suicide or hurting yourself in any way? If so, what caused it?
9	Share your recent experience of feeling down, depressed, or even desperate.
10	What are the problems or worries in your life? And how do you deal with them?
11	When do you feel like you are a terrible failure, or have let yourself or your family down?
12	When do you feel slower in your actions, thinking, or speaking?

Note: Sleep-related problems may be asked differently depending on contexts, such as How is your sleep recently? How long does it take to fall asleep? Will you wake up at night? Did you wake up too early? How long do you sleep every day? How easy it is to get a good quality of sleep?

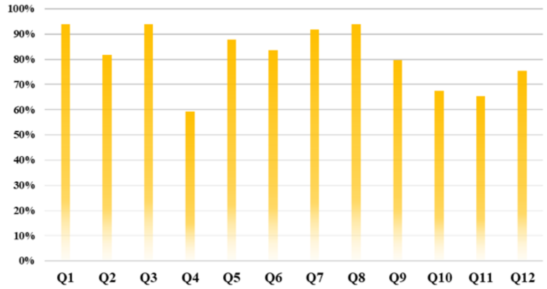


Fig. 4. Percentage of interview topics selected by clinicians which they would ask during diagnosis

and age biases are minimized to increase the statistical power. Once subjects met the inclusion criteria, they were instructed to proceed with the interview session.

Inclusion criteria of the MDD subjects were: (i) outpatient, age 18~65 years old, either sex; (ii) the diagnosis of MDD in accordance with DSM-IV as (a) established by the assessing psychiatrist, and (b) confirmed with M.I.N.I. 5.0.0; (iii) a HAMD-17 score >17 or a PHQ-9 score ≥ 9; and (iv) the informed consent was signed by the patient.

Exclusion criteria were: (i) any other psychiatric disorder or a mental disorder caused by a physical illness or substance abuse or a personality disorder; (ii) presence of psychotic symptoms during the depressive episodes; (iii) history of alcohol and substance dependence, acute poisoning; (iv) with serious risk of suicide, and (v) suffering from facial paralysis, disfigurement, unnatural facial movements, facial twitching, facial stroke, plastic surgery (with prosthesis), vitiligo, eye disease, speech disorder, stuttering, acoustic cord surgery, and other diseases.

3.2 Experiment Setup and Apparatus

Each interview was conducted by one of two research assistants (RAs) as interviewers. During the interview, only interviewers and interviewees were in the room (Fig. 3). Video and audio streams were captured separately. Based on the research literature on depression, we expect that the interpersonal nature of clinical interviews would improve the distinguishability across modalities [7]. A portable high-resolution audio recorder (SONY PCM-D100) was used to record subjects' voices and a digital camcorder (SONY HDR-CX680) for video recording. The camera is placed on a tripod behind the interviewer (in front of the interviewees). For each participant, the position of the camera was adjusted to ensure the optimal recording of facial areas. The audio recorder was placed on the desk at a distance of approximately 60cm. During the semi-structural interview, the interviewer and interviewee sat face to face. The audio was low pass filtered at 75 kHz, and the frame rate of video recording was 50 fps. All subjects were recorded using the same software and hardware apparatus. Though MDD and HC subjects were recorded at different sites, the room setting is the same (both are electromagnetic shielding rooms). Mobile phones were turned off or set to flight mode before recording. Participants were asked to avoid touching the recording stick and table during recording; prior to the key questions, the experimenter recorded for about 3 minutes before the official recording (did not involve the questions provided). All sessions were recorded during office hours. Due to the difficulty of MDD subject recruitment, the data were collected over a year.

3.3 Data Collection

The interview was semi-structured, starting from neutral issues, aiming to establish a harmonious relationship and make participants feel comfortable, then proceeded to key interview topics. The determination of interview topics is a key issue for the construction of a semi-structural interview based depression corpus. The simple way is to follow the established questionnaires such as HAMD and PHQ-9. However, this is quite straightforward since psychiatrists usually do not strictly follow any questionnaire during clinical practice. Considering the aim of this study, to build an automatic pre-screening tool through digital interviews with MDD patients that mimic the procedures of real clinicians, after a thorough discussion with senior clinicians, we developed a list of key interview questions based on the topics that clinicians may talk about during their practice, as well as questionnaires (PHQ-9, QIDS-SR [72], and HAMD). This process ended up with 35 questions by probing appetite, sleep, physical exercise, health condition, job/study, social interaction, memory, concentration, suicidal thoughts, family circumstances, etc., which would

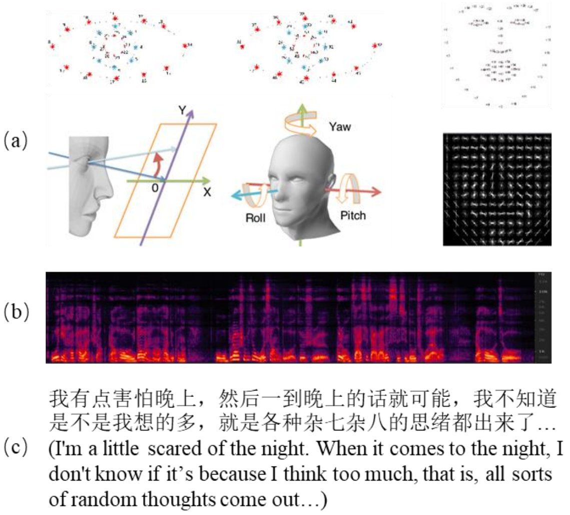
TABLE 4
DURATION AND WORD COUNT OF MDD AND HC SUBJECTSA INTERVIEW (IN MINUTES)

Part	MDD	HC	Total
Duration of full interviews:			
Total	783.25	1977.02	2760.27
Average	24.48	32.95	30.00
Standard deviation	5.91	3.97	6.22
Duration of subjects' speech:			
Total	201.19	450.32	651.52
Average	7.74	8.66	8.35
Standard deviation	4.57	2.81	3.52
Word count of subjects' speech:			
Total	33295	92197	125492
Average	1280	1773	1609
Standard deviation	609	920	861

TABLE 5
DESCRIPTION OF EXTRACTED VISUAL FEATURES

Feature sets	Description
Eye Gazes	Gaze direction in world coordinates, gaze angle in radians
Eye Landmarks	Eye landmarks in 2D (pixels), eye landmarks in 3D (millimeters)
Head Poses	Location of the head with reference to camera in millimeters, Rotation of head in radians.
Facial Landmarks	Facial landmarks in 2D (pixels), Facial landmarks in 3D (millimeters)
Facial Action Units	Intensities of 17 AUs (0 to 5), Presence of 18 AUs (0 absent, 1 present)

Note: see [2] *for a detailed explanation of extracted visual features.*



我有点害怕晚上，然后一到晚上的话就可能，我不知道是不是我想的多，就是各种杂七杂八的思绪都出来了…
(I'm a little scared of the night. When it comes to the night, I don't know if it's because I think too much, that is, all sorts of random thoughts come out…)

Fig. 5. Multimodal interview data. For each semi-structural interview: (a) visual features (2D&3D facial landmarks, eye landmarks, gaze directions, histogram of gradients, head rotations), (b) audio recording, visualized as spectrogram, and (c) transcription of the patient As speech (both in Chinese and English).

consume a large amount of time, and therefore not suitable for preliminary screening.

To shorten the topic list, we further surveyed key topics which may be asked during the clinical interview for diagnosis of MDD. Forty-nine MDD majored clinicians participated in this survey. All clinicians were first asked about the length of medical practice. Then they needed to select whether a certain question would be asked during their diagnosis. A total of 35 questions were displayed, and the questionnaire was ended with an extra question "Are there any other issues not displayed above that may be mentioned during your diagnosis? Please add".

The participated clinicians have a mean medical practice length of 9.4 years (SD = 7.0). After analysis of the survey, 12 questions, as shown in Table 3, were selected as key questions for the semi-structural interview. As shown in Fig. 4, the percentages of most key questions in clinical diagnosis and treatment are above 70%, except question 4 on friend communication which aims at probing social interactions. For the results of extra questions, there are 9 clinicians added questions. After a look into the extra questions, all the added questions were somehow covered in the question list. The selected 12 key questions were used in the interview session. Key interview topics were utilized to stimulate spontaneous speech, facial expressions, and related body gestures. The semi-structured interview was conducted by one of the trained RAs, and they were aware of the depression status of the subjects.

4 ANNOTATION & FEATURE EXTRACTION

4.1 Dialogue Segmentation

The audio and video recording of interviews included key topics about events and symptoms related to depression, as well as neutral questions designed to build rapport interactions. Only key interview questions were located and cropped from the original videos and audios with question information to support the development of AI tools for automatic screening. More specifically, we located the start and end of each key question by reviewing the whole audio stream and synchronizing each video clip simultaneously.

4.2 Interview Transcription

All interviews of key questions were further transcribed using software by iFlytek. Each transcription was reviewed and proofread for accuracy. The face-to-face interview was transcribed from the audio of the interviewee with labels of questions. A summary of annotated video and audio length is shown in Table 4. The total duration of the recorded interviews is over 46 hours. Moreover, the interviews were manually labeled to extract pure subjects' videos and audio of key questions. The total duration of pure speeches is about 651 minutes.

4.3 De-identification

All transcribed interviews were annotated to remove identifying information. Utterances that mention personal names, specific addresses, workplaces, and can be used to narrow the scope of the event were tagged and eliminated.

[2] https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format

Data annotated in the corpus does not contain protected health information. The facial signals are features, such as facial landmarks, eye landmarks, gaze angles, head movements, HOG features, AU intensity, and AU occurrence, which do not contain enough information for personal identification.

4.4 Visual Feature Extraction

MDD can be manifested through a variety of visual signs [73]. Such as changes in the activities of facial muscles, as well as eye gaze direction, often imply the persistent negative thoughts and feelings of sadness that characterize depression. Action Units (AUs) were proposed and defined by Ekman et al. [74] which describe the coordinated activities of facial muscle groups corresponding to specific facial expressions. AUs can be used to describe the changing characteristics of facial expression in depression since MDD patients often have poor expression ability. In addition, the occurrence of specific facial movements (smile, corner of mouth down, etc.) described by specific AUs (e.g. AU12) is directly related to depression. Thus, various studies have applied AUs to the automatic assessment of depression and achieved promising results [75]. Gaze angle, head pose, and facial landmarks have also been applied to depression assessment. The average gaze angle and change of gaze direction of the eye were also used as the detection characteristics of depression [76]. Facial landmarks can be applied in facial expression analysis. Therefore, AUs, eye gaze, head pose, and facial landmarks were extracted as visual features in the proposed depression corpus.

As mentioned in Section 3.1, 19 MDD and 26 HC have consented to video recording among all subjects. Automatic annotation of non-verbal features was carried out using an open-source facial behavior analysis toolkit called OpenFace (version 2.2.0) [77]. OpenFace is adopted since it is the state-of-art open-sourced tool for facial movement analysis, and is widely used in depression analysis, which may potentially facilitate corpus usage [34], [60]. The details of visual feature sets are shown in Table 5. The Openface features were frame-level but were averaged into sentence level with partitioned interview parts for subsequent statistical analysis and evaluation.

4.5 Acoustic Feature Extraction

Speech conveys non-verbal information on the depressive state of the speaker and the corresponding acoustic features are affected by depression [38]. For example, relevant studies have observed certain vocal changes in MDD patients, such as increased pause time and impaired fluency [40]. People with mental disorders also have shown disturbances in prosody [36]. Recently, machine learning with acoustic features has achieved a high-precision classification of suicidal thoughts in MDD patients [78]. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [79] is a quite standard feature set that is usually adopted as acoustic features in depression detection challenges such as AVEC [59]. Therefore, eGeMAPS were adopted as the acoustic feature set in the proposed corpus. Similar to visual features, a widely used open-source toolkit, OpenSMILE (version 3.0) [80], was used for acoustic feature extraction. Specifically, there are eight frequency-related parameters (pitch, jitter, frequency, and bandwidth of Formant 1, 2, 3), three energy/amplitude related parameters (loudness, shimmer, harmonics-to-noise ratio), as well as fourteen spectral parameters (alpha ratio, MFCC, etc.). See [79] for a detailed explanation of extracted features.

4.6 Textual Feature Extraction

As to textual features, semantic content from the interviewee's response can lead to a better estimation of the depression state [53]. There are several studies modeling depression from text-extracted features [42], [81], which have suggested textual features, such as the average word count, number of sentences, average number of words per sentence, and sentiment. These features are quite straightforward but have the advantage of their explainability. Similar to [42], We categorized textual features at word and sentence levels. We extracted these features from the text transcripts of interview responses. There are 6 word-level features and 4 sentence-level features. Word-level features are ratios of adjectives, adverbs, exclamations, verbs, modal particles, and the total number of words count. Sentence-level features are ratios of positive sentences, negative sentences, the sentiment of the whole response, and the number of sentences. Chinese word cut, part-of-speech tagging as well as sentiment analysis were achieved with Xmnlp (version 0.3.1) [82] which is an open-source lightweight Chinese natural language processing toolkit.

4.7 Visualization of the Corpus

An illustration of multimodal interview features and data is shown in Fig. 5. In order to have a better insight into the data distributions, we used the t-distributed stochastic neighbor embedding (t-SNE) [83], which is a nonlinear dimensionality reduction technique, to visualize the corpus with features of various modalities. As shown in Fig. 6, one can observe that the multimodal features seem to be better clustered than the unimodal features in terms of two class labels. The unimodal features, especially visual and textual features, are not very discriminative even though statistical analyses show significant differences between several of the extracted features. These are intuitive motivations of machine/deep learning tools and multimodal fusion methods may aid the classification process with improved performance in the proposed corpus.

4.8 Deep Representations

The above extracted features have advantages of their interpretability when analyzing behavioral patterns between MDD and HC. Besides that, features extracted with deep learning models pertrained on large scale datasets may have more powerful representation abilities which may benefit subsequent machine learning tasks. Therefore, we also extracted a deep representation of visual features based on a newly developed transformer model, TimesFormer [84], pretrained on a large scale video understanding task (Kinetics-600 [85]). The dimension of extracted feature is 768 for each question-level video clip and

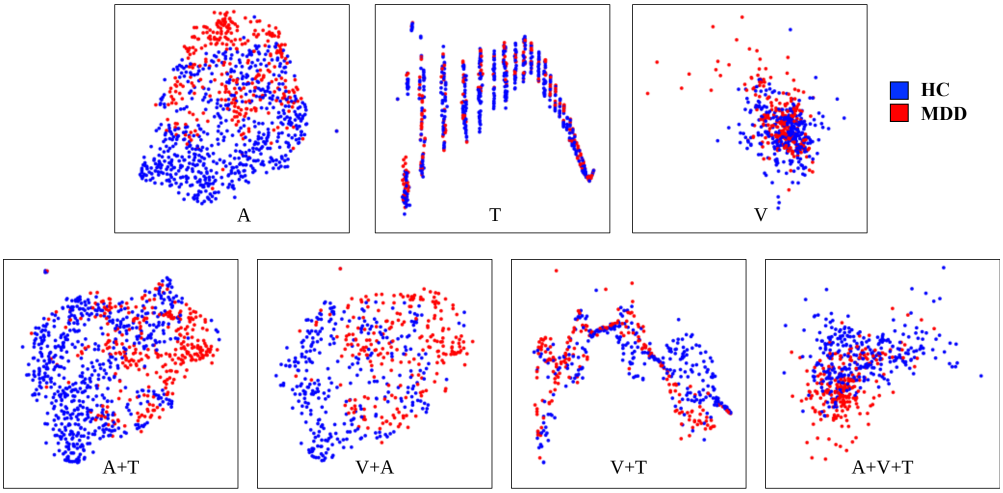


Fig. 6. T-SNE 2D visualization, each dot represents one partitioned interview section (For modality of A, T, and A+T, results were obtained from data of all 78 participants. As to V and V related features, results were from the 45 participants with video recordings).

the model URL is given in [3]. Timesformer is a state-of-art model published in 2021 and was adopted because of its good performance and availability. Though Kinetics-600 is not emotion-focused, a pretrained model on this large scale video dataset may have an advantage in temporal motion modeling. Furthermore, for deep representation of audio, a VGG-like audio classification model, VGGish, pretrained on a preliminary version of YouTube-8M was adopted [82]. This model outputted 128-dimensional embeddings for each question-level audio. And the textual feature was also extracted with Chinese BERT [79], pretrained weights available at [4], which is a transformer-based text embedding model that achieved state-of-art performance on a series of natural language understanding tasks. We obtained 768 dimension embedding for the text of each partitioned interview.

The deep representation, along with the Openface and eGeMAPS features, are shared together in the download link provided, which hopefully makes the proposed corpus more open to further research.

5 Statistical Analysis & Benchmark Evaluation

5.1 Statistical Analysis of Visual, Acoustic, and Textual Features

To compare the differences in visual, acoustic, and textual modalities between MDD and HC groups, multiple analysis of covariance (MANCOVA) was used for statistical analysis. Partial square of Eta (η_p^2) was reported in the analyses of effect sizes. η_p^2 values of 0.01, 0.06, and 0.14 were considered as small, moderate, and large effect sizes respectively [35], [86].

5.1.1 Visual feature analysis

A total of 39 visual features were analyzed for videos of each interview question per subject. There were 4 gaze-related features, (mean and standard deviation of horizontal and vertical gaze angles, SD represents an estimate of the variance of gaze angle), mean intensity of 17 AUs, the occurrence rate of 18 AUs, see [77] for a detailed explanation of features. A three-way MANCOVA analysis was conducted to test for the main effects of subject group, interview question, and gender. The results reveal the main effects of group (Wilks' Lambda $F(49, 436) = 18.14$, $p<.001$, $\eta_p^2 = 0.67$) and gender (Wilks' Lambda $F(49, 436) = 9.54$, $p<.001$, $\eta_p^2=0.52$) as well as an interaction effect of group and gender (Wilks' Lambda $F(49, 436) = 9.72$, $p<.001$, $\eta_p^2=0.52$). This indicates salient visual differences between groups and genders and the significant visual features of MDD vary with gender. For individual visual features, however, none of them has a large effect size. Only occurrence rate of AU06_c (Cheek Raiser, $F(1, 530)=14.36$, $p<.001$, $\eta_p^2=.098$) has a significant difference with a medium effect. The average intensity of AU01 (Inner Brow Raiser), AU02 (Outer Brow Raiser), AU04 (Brow Lowerer), AU12 (Lip Corner Puller), AU20 (Lip stretcher), AU23 (Lip Tightener), and AU25 (Lips part), as well as the occurrence rate of AU04 (Brow Lowerer), AU12 (Lip Corner Puller), AU14 (Dimpler), AU17 (Chin Raiser), AU20 (Lip stretcher), AU23 (Lip Tightener), AU25 (Lips part), AU28 (Lip Suck) were found differ significantly between MDD and HC group with a small effect size. Although gaze-related parameters do not show a large effect size of significant difference, vertical gaze angle has a p-value less than 0.001, which describes the more downward angle of gaze for MDD.

These results are consistent with previous findings of [23], [87], which found reduced affiliative expressions and increased non-affiliative expressions in MDD. In particular, both studies found the higher average intensity of AU01, AU04 which is associated with sadness emotion, and lower intensity of AU06 which is a common expression for happiness [75]. All significant AUs are around the eye and mouth areas. The detailed statistics of visual features between MDD and HC are given in Table S1 of the supplementary materials.

[3] https://www.drobox.com/s/4h2qt41m2z3aqrb/TimeSformer_divST_8 x32_224_K600.pyth?dl=0

[4] https://github.com/ymcui/Chinese-BERT-wwm

5.1.2 Acoustic feature analysis

For acoustic features, a three-way MANCOVA analysis reveals the main effects of group (Wilks' Lambda F(25, 845) = 68.14， p<.001, η_p^2=0.67) and gender (Wilks' Lambda F(25, 845) =29.78， p<.001, η_p^2=0.47) as well as an interaction effect of group and gender (Wilks' Lambda F(25, 845) = 14.30， p<.001, η_p^2=0.30). This indicates salient acoustic parameters differences between groups and genders, and the significant acoustic features of MDD vary with gender. For individual acoustic parameters, only ones with large effect sizes were reported as significant features. Although main effects of subject group were found on most acoustic parameters (except mfcc1, Harmonic difference H1-H2, H1-A3, and Alpha Ratio), only mfcc4 (F(1, 915)=245.22, p<.001， η_p^2 =0.21), F1_Bandwidth (F(1, 915)=272.79, p<.001 ， η_p^2 =0.23), F2_Bandwidth (F(1, 915)=333.86, p<.001 ， η_p^2 =0.27), Hammarberg Index (F(1, 915)=430.63, p<.001, η_p^2=.32) Spectral Slope 0-500 Hz (F(1, 915)=920.10, p<.001, η_p^2=.50) were significantly different between two groups (MDD and HC) with large effect sizes. To further assess the effects of question and gender on the five significant features, the MANCOVA results of the five features were analyzed. No main effect of question and gender, as well as the interaction effect between question and group, gender and group, and question, gender, and group was found (all η_p^2s<0.14).

Loudness represents the estimate of perceived signal intensity and has been studied intensively in the detection of depression[81], [88]. The results of this study show that pitch and loudness are both lower in MDD group, which are in agreement with previous studies [35] and clinical observation [89] that MDD subjects are generally believed to have a lower sound volume than HC subject. MFCC is Mel-Frequency Cepstral Coefficients that represent acoustic tract changes [35]. Hammarberg Index is the ratio of the highest energy peak of 0-2 kHz region to that of 2–5 kHz region. The spectral slope is the linear regression slope of the logarithmic power spectrum within two given frequency bands [79]. Both Hammarberg Index and spectral slopes are previously reported to have a significant correlation with depression severity [90]. Significant differences in formant 1, 2 bandwidths were also found between the two groups. Formant features have already been shown to capture information useful in distinguishing between the two classes [39] and have also been linked with emotional and cognitive information. The detailed statistics of 25 acoustic features between MDD and HC groups are shown in Table S2 of the supplementary materials.

5.1.3 Textual feature analysis

As to textual features, a three-way MANCOVA analysis reveals main effect of subject group (Wilks' Lambda F(10, 863)= 14.90， p=.001, η_p^2 =0.018) and interview question (Wilks' Lambda F(110, 6471.45) = 1.46， p<.001, η_p^2=0.147), no significant effects of gender and interactions among them (p=.11). MDD subjects tend to talk less and use fewer verbs than HC group, and their sentences have a higher ratio of negative sentiments, which are consistent with clinical observations. The detailed statistics of 10 textual features between MDD and HC groups are shown in Table S3 of the supplementary materials.

5.2 Benchmark Evaluation

Benchmark evaluations of the proposed dataset were also conducted to provide a basis for any follow-up multimodal depression assessment research. The clinical diagnosis of MDD was considered as the ground truth for the classification task and the scores of PHQ-9 questionnaire were the ground truth for the regression task. Previously extracted features of audio, video, and text modalities (A, V, T) and their combinations (A+V, A+T, V+T, A+V+T) were tested for depression assessment. For each subject, there are a total of 888 features (25*12 acoustic features, 10*12 textual features, and 39*12 visual features if video recorded, where 12 is the number of questions).

5.2.1 Classification task

Precision, recall, F-measures, and AUROC are reported as evaluation metrics. Linear kernel support vector machine

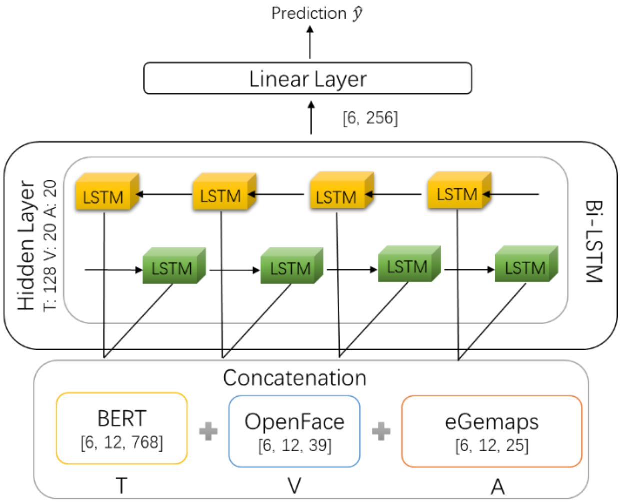


Fig. 7. Early Fusion Bi-directional LSTM for depression classification and regression.

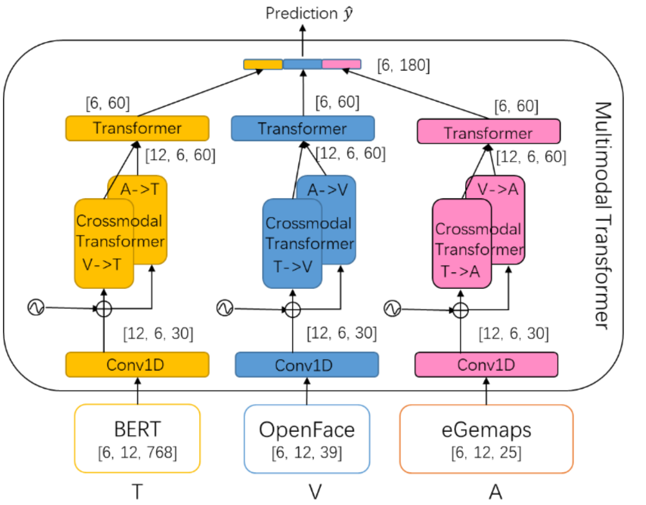


Fig. 8. Multimodal Transformer-based fusion for depression classification and regression.

(SVM) [91], SVM based on sequential minimal optimization (SMO) [92], Logistic regression, and Naïve Bayes were adopted as the baseline classifier with stratified 5-fold cross-validation, all were implemented with default parameters in Weka [93]. The results are listed in Table 6, better performance was achieved with acoustic and textual features, both A+T and A+V+T. Facial features seem not very discriminative between two groups even with a quite large number of features.

Furthermore, we also evaluated two more deep learning based methods on the proposed dataset for both depression classification and regression as shown in Fig.7 and Fig. 8. The first one is an early fusion bi-directional LSTM (EF-Bi-LSTM) [94] model which takes the concatenation of textual, visual, and acoustic features as input. The visual and acoustic features were extracted with Openface and OpenSmile, respectively, same as other baseline methods. But the textual feature is encoded with pretrained Chinese BERT [95] embedding. Instead of early fusion, the second one is based on more advanced multimodal transformer fusion [57] which can handle inherent data non-alignment and long-range dependencies across modalities. The layers and parameters settings are illustrated in the two figures. For reproducibility, we share the code of Bi-LSTM and MuIT methods on the GitHub website: https://github.com/CMDC-corpus/CMDC-Baseline. For both models, we used Adam [96] to optimize with a batch size of 6, and a learning rate of 1e^{-4}. To overcome overfitting, dropout layers were inserted in Bi-LSTM and transformer layers with a dropout rate of 0.1. The number of hidden units for A, V, T are 20, 20, 128 respectively and the number of training epochs was set as 200. The parameters were chosen accordingly for different fusion methods (refer to the shared code for details). As shown in the last two columns of Table 6, the (EF) Bi-LSTM model performs comparably to other baseline methods. But the multimodal transformer fusion model generally performs better than other fusion methods. As shown in the above results, the depression classification performance almost reaches the bottleneck with an F1 score of 0.95. Therefore, We further conducted a more difficult regression task.

5.2.2 Regression task

For the regression task, mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficients were reported as evaluation metrics. The mean PHQ-9 score computed on the training set for prediction was conducted as a trivial baseline. A MAE of 7.03 and a RMSE of 7.84 are the trivial baseline results for 45 subjects with all modalities,. Linear regression, Support vector regression, Multi layer perceptron, and K nearest neighborhood were also adopted as the baseline regressors with stratified 5-fold cross-validation, all were implemented with default parameters in Weka. As listed in Table 7. Some of the machine learning methods are even worse than the trivial baseline, but the two deep learning models always outperform it. The best result is achieved by multimodal transformer fusion of acoustic and textual modalities with MAE 3.66, RMSE 4.59, and Pearson 0.72. The fu-

TABLE 6
BENCHMARK EVALUATION OF DEPRESSION CLASSIFICATION WITH VARIOUS MODALITIES

	Metric	SVM (Linear)	SVM (SMO)	Logistic	Naïve Bayes	Bi-LSTM (Early Fusion)	MulT
A	Precision	0.78	*0.92*	0.85	0.89	**1.00**	—
	Recall	0.78	**0.91**	0.84	*0.89*	0.83	—
	F1	0.78	**0.91**	0.84	0.89	**0.91**	—
	AUROC	0.77	*0.90*	0.77	0.89	**0.92**	—
T	Precision	0.71	0.84	**0.93**	0.85	*0.87*	—
	Recall	0.71	0.84	**0.93**	0.80	*0.90*	—
	F1	0.71	0.84	**0.93**	0.78	*0.88*	—
	AUROC	0.71	*0.84*	**0.99**	*0.84*	0.83	—
V	Precision	0.61	0.69	*0.73*	0.60	**1.00**	—
	Recall	0.60	0.69	**0.73**	0.60	*0.71*	—
	F1	0.60	0.69	*0.73*	0.60	**0.83**	—
	AUROC	0.60	0.67	*0.70*	0.57	**0.86**	—
A+T	Precision	0.90	*0.92*	*0.92*	0.91	**0.97**	0.87
	Recall	*0.91*	*0.91*	*0.91*	0.89	*0.91*	**0.96**
	F1	*0.91*	*0.91*	*0.91*	0.89	**0.94**	*0.91*
	AUROC	*0.90*	*0.90*	0.89	0.86	**0.91**	0.87
V+A	Precision	0.78	**0.89**	0.80	0.82	0.83	*0.87*
	Recall	0.78	**0.87**	0.80	0.82	*0.83*	**0.87**
	F1	0.78	*0.86*	0.80	0.82	0.83	**0.87**
	AUROC	0.77	*0.84*	0.74	0.80	0.83	**0.87**
V+T	Precision	0.72	*0.87*	0.80	0.66	0.82	**1.00**
	Recall	0.71	*0.87*	0.80	0.67	**0.89**	0.83
	F1	0.71	*0.87*	0.80	0.66	0.85	**0.91**
	AUROC	0.71	*0.86*	0.78	0.69	0.82	**0.91**
A+V+T	Precision	*0.91*	*0.91*	0.82	0.84	0.87	**0.97**
	Recall	**0.91**	*0.89*	0.82	0.84	*0.89*	0.85
	F1	**0.91**	0.89	0.82	0.84	0.88	**0.91**
	AUROC	**0.90**	0.87	0.82	0.81	0.82	*0.88*

The best results are in Bold and the second-best results are in Italic.

TABLE 7
BENCHMARK EVALUATION OF DEPRESSION REGRESSION WITH VARIOUS MODALITIES

	Metric	Linear Regression	SVR (SMOreg)	MLP	KNN	Bi-LSTM (Early Fusion)	MulT
A	MAE	8.10	7.98	6.26	*5.78*	**4.59**	—
	RMSE	11.6	11.93	8.83	*7.94*	**6.14**	—
	Pearson	0.48	*0.53*	0.48	0.46	**0.64**	—
T	MAE	6.77	6.65	*4.77*	5.96	**3.81**	—
	RMSE	10.25	10.58	*6.47*	8.08	**5.35**	—
	Pearson	0.48	0.50	**0.64**	0.59	*0.63*	—
V	MAE	7.09	6.81	*6.73*	9.13	**5.77**	—
	RMSE	9.57	9.50	*9.02*	11.79	**7.70**	—
	Pearson	*0.44*	**0.47**	0.33	-0.05	0.28	—
A+T	MAE	7.00	6.10	5.34	5.27	*3.68*	**3.66**
	RMSE	10.26	8.71	7.12	7.15	*4.62*	**4.59**
	Pearson	0.38	0.48	0.53	0.60	*0.71*	**0.72**
V+A	MAE	6.78	6.15	7.09	**5.04**	5.18	*5.08*
	RMSE	9.74	8.70	12.63	7.66	*6.43*	**6.02**
	Pearson	0.40	0.45	0.35	*0.57*	0.55	**0.60**
V+T	MAE	5.41	5.13	17.83	7.15	*4.76*	**4.61**
	RMSE	6.74	6.28	83.59	10.03	*5.75*	**5.56**
	Pearson	0.62	**0.67**	0.19	0.22	0.54	*0.64*
A+V+T	MAE	5.80	5.31	5.89	5.89	*4.55*	**4.32**
	RMSE	7.97	7.04	7.60	8.38	*5.67*	**5.61**
	Pearson	0.50	0.57	0.39	0.50	*0.68*	**0.72**

The best results are in Bold and the second-best results are in Italic.

TABLE 8
EVALUATIONS ON DAIC

		Precision	Recall	F1	MAE	RMSE
A	[16]*	0.71	0.56	0.63	5.13	6.50
	BiLSTM	0.88	0.70	0.77	5.20	6.51
T	[16]*	0.57	0.80	0.67	5.18	6.38
	[41]*	0.89	0.85	0.87	4.15	5.51
	BiLSTM	0.91	0.81	0.86	4.41	5.36
V	[98]*	0.67	0.91	0.78	5.01	6.32
	[99]*	—	—	—	4.61	5.78
	BiLSTM	0.94	0.74	0.83	5.22	6.40
A + T	[16]*	0.71	0.83	0.77	5.10	6.37
	[100]*	0.79	0.92	0.85	3.75	5.44
	BiLSTM	0.91	0.83	0.87	4.82	5.97
	MulT	0.88	0.81	0.84	4.74	5.81
V + A	[20]	—	—	—	4.20	5.51
	[101]	—	—	—	5.39	6.34
	BiLSTM	0.91	0.73	0.81	5.11	6.38
	MulT	0.91	0.75	0.82	4.67	5.88
V + T	BiLSTM	0.91	0.86	0.88	4.64	5.78
	MulT	0.97	0.84	0.9	4.28	5.47
A + V + T	[30]*	0.71	0.83	0.77	3.67	—
	[102]	0.80	—	0.81	3.61	4.99
	BiLSTM	0.94	0.79	0.86	4.77	5.97
	MulT	0.94	0.81	0.87	4.64	5.77

** denotes results on the validation set.*

sion of three modalities also achieved relatively good results. Similar to the classification task, the fusion of multimodal generally perform better than the unimodal methods.

For further direction, more advanced representation learning techniques for feature extraction, embedding, and multimodal fusion techniques are worth exploring for better performance [97].

5.2.3 Evaluation of baseline methods on DAIC

To further analyze the benchmark evaluation, we also evaluated the two baseline deep-learning methods on the publicly available English depression dataset DAIC [58]. We implemented two deep learning models for both classification and regression tasks with the same feature extraction of V and T (with pretrained English Bert model) modalities. For audio, the DAIC provides 74 audio features extracted with the COVAREP toolbox (v. 1.3.2) [103] on every 10ms segment. The mean, max, and min of features on partitioned question-level audio were calculated for temporal aggregation. The training set and the validation set of DAIC were combined for training, and the results were obtained on the test set. Since DAIC is widely used in the literature, we also listed some previous results as a comparison. As shown in Table 8, the performance of the two methods are close to the previous state-of-art. By comparing Table 8 with Tables 6 and 7, we can observe that both the classification and regression performance on the proposed CMDC are comparable but slightly better than that on the DAIC, which may be due to the well-defined semi-structural interview. Considering the classification accuracy on the proposed corpus, the feasibility of preliminary screening based on multimodal behavior signals thus could somehow be guaranteed.

6 DISCUSSION & CONCLUSION

Previous studies found that depressed patients differ from the HC group in observable behavior modalities including visual, acoustic, and verbal signals [27], [34], [35], [38], [88]. Such as visual signals from the facial area show that psychological distress is predicted by larger downward angles of gaze, shorter average duration, and less intensity of smiles [33]. Depressed speech has found several distinguishing acoustic features as indicators of disease severity and treatment efficacy [104]. Therefore, a depression corpus with multidimensional behavior observations would promote the research of automatic depression detection enabled by machine learning methods for the fusion of multimodal features, which could collectively provide a stronger indication for depression.

In this paper, we proposed a semi-structural interview based Chinese multimodal depression corpus with clinically validated depressive subjects. The proposed dataset can be helpful to explore objective indicators of MDD in multiple behavior modalities, and these indicators are planned to be implemented in a virtual AI agent for MDD preliminary screening (as shown in Fig. 1), to allow the identification of people who should be referred to further evaluation. The release of this dataset would promote the research progress and practical applications of depression screening and assessment. And in a broader context, to show how core affective computing research is becoming an important driving factor in the application research of solving societal problems of mental health.

Compared with existing depression datasets, this corpus has the following merits:

- clinical diagnosed pure MDD patients as subjects, unlike the other datasets which also contain PTSD, anxiety, and Melancholia subjects;
- extensive survey of MDD majored clinicians for determination of key interview questions to best mimic real diagnosis scenarios;
- both experts assessed HAMD and self-reported PHQ-9 are available as ground truth scores for the development of automatic screening tools;
- publicly available multimodal Chinese depression dataset based on clinically validated semi-structural interviews.

Limitation and future work. Due to the rigorous selection criteria, one limitation is the relatively modest number of subjects which is common in similar studies. A larger-scale study is preferable, however, due to the difficulty of MDD subject recruitment and lack of clinical assessment, especially during the COVID-19 pandemic, it is particularly hard to construct a larger dataset. Moreover, the poor regression results show that the assessment of the severity of depression is still a challenge. One reason for the high classification accuracy may due to that although our experimental control is semi-structured, it is still relatively strict, which is a double-edged sword. For example, the acquisition environment is an electromagnetic shielding room, there is no noise from natural scenes, and the text transcription uses proofreading after the automatic transcription tools. Classification and regression based on automatic transcripted text (without proofreading) are worthy of further exploration. Besides, speech annotation and speaker separation were manually achieved in this work, which may be processed automatically with advanced speech separation techniques. The corpus released still has a large room for improvement in the unimodal depression detection task based on vision and text modalities. A fully automated system to analyze multimodal signals for depression detection was not the main focus of this study. In future work, we plan to further investigate the discriminative dynamics, acoustic prosody, and text semantics individually and in combination for depression screening. This study shows the feasibility of the preliminary screening of depression with semi-structured interviews. We hope this corpus could promote the research progress and practical applications of depression screening and assessment with core affective computing technologies.

Distribution. The dataset is shared upon request for research purposes. The de-identified portions of the data are intended to be made more widely available to the research community.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. U20B2062, U19B2032). The authors would like to thank Guangyao Sun, Jingyu Hou, Xuesong Wang, and Yubo An for their generous assistance during data collection and processing.

REFERENCES

[1] A. T. Beck and B. A. Alford, *Depression: Causes and treatment*. University of Pennsylvania Press, 2009.

[2] World Health Organization, "Depression and other common mental disorders: global health estimates," World Health Organization, 2017.

[3] Y. Huang *et al.*, "Prevalence of mental disorders in China: a cross-sectional epidemiological study," *The Lancet Psychiatry*, vol. 6, no. 3, pp. 211–224, 2019.

[4] Y. Chen *et al.*, "Patterns and correlates of major depression in Chinese adults: a cross-sectional study of 0.5 million men and women," *Psychol. Med.*, vol. 47, no. 5, pp. 958–970, 2017.

[5] W. Li *et al.*, "The first national action plan on depression in China: Progress and challenges," *Lancet Reg. Heal. Pacific*, vol. 7, 2021.

[6] S. Graham *et al.*, "Artificial Intelligence for Mental Health and Mental Illnesses: an Overview," *Curr. Psychiatry Rep.*, vol. 21, no. 11, 2019, doi: 10.1007/s11920-019-1094-0.

[7] H. Dibeklioglu, Z. Hammal, and J. F. Cohn, "Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding," vol. 22, no. 2, pp. 525–536, 2018.

[8] S. Gao, V. D. Calhoun, and J. Sui, "Machine learning in major depression: From classification to treatment outcome prediction," *CNS Neurosci. Ther.*, vol. 24, no. 11, pp. 1037–1052, 2018.

[9] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation," no. 2, pp. 37–43, 2017, doi: 10.1145/3133944.3133947.

[10] A. P. Association, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[11] A. F. Schatzberg, "Scientific issues relevant to improving the diagnosis, risk assessment, and treatment of major depression," *Am. J. Psychiatry*, vol. 176, no. 5, pp. 342–347, 2019, doi: 10.1176/appi.ajp.2019.19030273.

[12] Y. Liang, X. Zheng, and D. D. Zeng, "A Survey on Big Data-Driven Digital Phenotyping of Mental Health," *Inf. Fusion*, 2019, doi: 10.1016/j.inffus.2019.04.001.

[13] M. A. X. Hamilton, "Development of a rating scale for primary depressive illness," *Br. J. Soc. Clin. Psychol.*, vol. 6, no. 4, pp. 278–296, 1967.

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

[15] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing : From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017, doi: 10.1016/j.inffus.2017.02.003.

[16] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, pp. 1716–1720, 2018, doi: 10.21437/Interspeech.2018-2522.

[17] W. Xie *et al.*, "Interpreting Depression From Question-wise Long-term Video Recording of SDS Evaluation," *IEEE J. Biomed. Heal. Informatics*, vol. PP, no. 8, p. 1, 2021, doi: 10.1109/JBHI.2021.3092628.

[18] P. Zhang, M. Wu, H. Dinkel, and K. Yu, *DEPA: Self-Supervised Audio Embedding for Depression Detection Pingyue*, vol. 1, no. 1. Association for Computing Machinery, 2021.

[19] E. Toto, M. L. Tlachac, and E. A. Rundensteiner, "AudiBERT: A Deep Transfer Learning Multimodal Classification Framework for Depression Screening," in *Conference on information and knowledge Management*, 2021, vol. 1, no. 1, pp. 4145–4154, doi: 10.1145/3459637.3481895.

[20] Z. Zhao *et al.*, "Automatic Assessment of Depression from Speech via a Hierarchical Attention Transfer Network and Attention Autoencoders," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 423–434, 2020, doi: 10.1109/JSTSP.2019.2955012.

[21] Y. Zhang, Y. Wang, X. Wang, B. Zou, and H. Xie, "Text-based Decision Fusion Model for Detecting Depression," *ACM Int. Conf. Proceeding Ser.*, pp. 101–106, 2020, doi: 10.1145/3421515.3421516.

[22] X. Sun, Y. Song, and M. Wang, "Toward Sensing Emotions with Deep Visual Analysis: A Long-Term Psychological Modeling Approach," *IEEE Multimed.*, vol. 27, no. 4, pp. 18–27, 2020, doi: 10.1109/MMUL.2020.3025161.

[23] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral Representation of Behaviour Primitives for Depression Analysis," *IEEE Trans. Affect. Comput.*, vol. 14, no. 8, pp. 1–16, 2020, doi: 10.1109/TAFFC.2020.2970712.

[24] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 578–584, 2018, doi: 10.1109/TAFFC.2017.2650899.

[25] S. Gao and V. D. Calhoun, "Machine learning in major depression : From classification to treatment outcome prediction," no. April, pp. 1037–1052, 2018, doi: 10.1111/cns.13048.

[26] W. Carneirodemelo, E. G. Granger, and M. Bordallo Lopez, "MDN: A Deep Maximization-Differentiation Network for Spatio-Temporal Depression Detection," *IEEE Trans. Affect. Comput.*, vol. XX, no. X, pp. 1–1, 2021, doi: 10.1109/taffc.2021.3072579.

[27] S. M. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, and G. Parker, "Interpretation of Depression Detection Models via Feature Selection Methods," *IEEE Trans. Affect. Comput.*, vol. X, no. X, pp. 1–18, 2020, doi: 10.1109/TAFFC.2020.3035535.

[28] S. Alghowinem *et al.*, "Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 478–490, 2018, doi: 10.1109/TAFFC.2016.2634527.

[29] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, "Multimodal assessment of depression from behavioral signals," in *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, 2018, pp. 375–417.

[30] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, "Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions," pp. 1–7, 2018, [Online]. Available: http://arxiv.org/abs/1811.08592.

[31] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image Vis. Comput.*, vol. 32, no. 10, pp. 641–647, 2014, doi: 10.1016/j.imavis.2013.12.007.

[32] N. Carvalho *et al.*, "Eye movement in unipolar and bipolar depression: A systematic review of the literature," *Front. Psychol.*, vol. 6, no. DEC, 2015, doi: 10.3389/fpsyg.2015.01809.

[33] S. Scherer *et al.*, "Automatic behavior descriptors for psychological disorder analysis," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[34] A. Pampouchidou *et al.*, "Automatic Assessment of Depression Based on Visual Cues: A Systematic Review," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 445–470, 2019, doi: 10.1109/TAFFC.2017.2724035.

[35] J. Wang, L. Zhang, T. Liu, W. Pan, B. Hu, and T. Zhu, "Acoustic differences between healthy and depressed people: A cross-situation study," *BMC Psychiatry*, vol. 19, no. 1, 2019, doi: 10.1186/s12888-019-2300-7.

[36] L. Zhang, R. Duvvuri, K. K. L. Chandra, T. Nguyen, and R. H. Ghomi, "Automated voice biomarkers for depression symptoms using an online cross ‐ sectional data collection initiative," *Depress. Anxiety*, vol. 37, no. 7, pp. 657–669, 2020.

[37] J. R. Williamson *et al.*, "Detecting Depression using Vocal , Facial and Semantic Communication Cues," pp. 11–18, 2016.

[38] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, 2015, doi: 10.1016/j.specom.2015.03.004.

[39] D. J. France and R. G. Shiavi, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, 2000, doi: 10.1109/10.846676.

[40] D. Raucher-Chéné, A. M. Achim, A. Kaladjian, and C. Besche-Richard, "Verbal fluency in bipolar disorders: A systematic review and meta-analysis," *J. Affect. Disord.*, vol. 207, pp. 359–366, 2017.

[41] H. Dinkel, M. Wu, and K. Yu, "Text-based Depression Detection: What Triggers An Alert," 2019, [Online]. Available: http://arxiv.org/abs/1904.05154.

[42] S. Guohou, Z. Lina, and Z. Dongsong, "What reveals about depression level? The role of multimodal features at the level of interview questions," *Inf. Manag.*, vol. 57, no. 7, p. 103349, 2020, doi: 10.1016/j.im.2020.103349.

[43] Y. Gong and C. Poellabauer, "Topic Modeling Based Multi-modal Depression Detection," *Proc. 7th Annu. Work. Audio/Visual Emot. Chall. - AVEC '17*, pp. 69–76, 2017, doi: 10.1145/3133944.3133945.

[44] L. Zhang, R. Duvvuri, K. K. L. Chandra, T. Nguyen, and R. H. Ghomi, "Automated voice biomarkers for depression symptoms using an online cross ‐ sectional data collection initiative," *Depress. Anxiety*, vol. 37, no. 7Kenton, M. C., Kristina, L., Devlin, J. (2017). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding., pp. 657–669, 2020.

[45] W. Carneirodemelo, E. G. Granger, and M. Bordallo Lopez, "MDN: A Deep Maximization-Differentiation Network for Spatio-Temporal Depression Detection," *IEEE Trans. Affect. Comput.*, no. April, pp. 1–1, 2021, doi: 10.1109/taffc.2021.3072579.

[46] S. Hong, A. Cohn, D. C. Hogg, and E. With, "Using Graph Representation Learning with Schema Encoders to Measure the Severity of Depressive Symptoms," in *International Conference on Learning Representations*, 2021, pp. 1–23.

[47] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 53–59.

[48] J. Xu, S. Song, K. Kusumam, H. Gunes, and M. Valstar, "Two-stage Temporal Modelling Framework for Video-based Depression Recognition using Graph Representation," pp. 1–16, 2021, [Online]. Available: http://arxiv.org/abs/2111.15266.

[49] F. Ringeval, M. Valstar, N. Cummins, R. Cowie, and M. Schmitt, "AVEC 2019 Workshop and Challenge: State-of-Mind, Depression with AI, and Cross-Cultural Affect Recognition," 2019.

[50] A. Ray, "Multi-level Attention Network using Text , Audio

and Video for Depression Prediction," pp. 81–88, 2019.

[51] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 2017, pp. 131–135.

[52] L. He *et al.*, "Deep learning for depression recognition with audiovisual cues: A review," *Inf. Fusion*, vol. 80, pp. 56–86, 2022.

[53] J. R. Williamson *et al.*, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 11–18.

[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv Prepr. arXiv1810.04805*, 2018.

[55] M. Valstar *et al.*, "AVEC 2016 - Depression, mood, and emotion recognition workshop and challenge," *AVEC 2016 - Proc. 6th Int. Work. Audio/Visual Emot. Challenge, co-located with ACM Multimed. 2016*, pp. 3–10, 2016, doi: 10.1145/2988257.2988258.

[56] F. Ringeval *et al.*, "AVEC 2017 – Real-life Depression, and Affect Recognition Workshop and Challeng," *Proc. 7th Annu. Work. Audio/Visual Emot. Chall. - AVEC '17*, pp. 3–9, 2017, doi: 10.1145/3133944.3133953.

[57] Y. H. H. Tsai, S. Bai, P. P. Liang, J. Zico Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 6558–6569, 2020, doi: 10.18653/v1/p19-1656.

[58] J. Gratch *et al.*, "The Distress Analysis Interview Corpus of human and computer interviews," *Lrec*, pp. 3123–3128, 2014, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.495.3966&rep=rep1&type=pdf.

[59] F. Ringeval *et al.*, "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition," no. Avec, 2019, [Online]. Available: http://arxiv.org/abs/1907.11510.

[60] W. Lin, I. Orton, Q. Li, G. Pavarini, and M. Mahmoud, "Looking At The Body: Automatic Analysis of Body Gestures and Self-Adaptors in Psychological Distress," *IEEE Trans. Affect. Comput.*, vol. 14, no. 8, 2021, doi: 10.1109/TAFFC.2021.3101698.

[61] 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, 2021, doi: 10.1109/TBME.2020.3010472.

[62] W. Guo, H. Yang, Z. Liu, Y. Xu, and B. Hu, "Deep Neural Networks for Depression Recognition Based on 2D and 3D Facial Expressions Under Emotional Stimulus Tasks," *Front. Neurosci.*, vol. 15, no. April, 2021, doi: 10.3389/fnins.2021.609760.

[63] H. Cai, S. Shuting, F. Tian, and H. Xiao, "MODMA dataset : a Multi-model Open Dataset for Mental- disorder Analysis Background & Summary," no. March, 2020.

[64] Y. Shen, H. Yang, and L. Lin, "Automatic Depression Detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based Model," *arXiv Prepr. arXiv2202.08210*, 2022.

[65] 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. Heal. Informatics*, vol. 24, no. 3, pp. 815–824, 2020, doi: 10.1109/JBHI.2019.2930604.

[66] K. E. B. Ooi, L.-S. A. Low, M. Lech, and N. Allen, "Prediction of clinical depression in adolescents using facial image analysis," *Image Anal. Multimed. Interact. Serv. WIAMIS*, vol. 10, 2011.

[67] K. Y. Huang, C. H. Wu, Y. T. Kuo, and F. L. Jang, "Unipolar depression vs. bipolar disorder: An elicitation-based approach to short-term detection of mood disorder," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, no. October 2017, pp. 1452–1456, 2016, doi: 10.21437/Interspeech.2016-620.

[68] W. W. K. Zung, "A self-rating depression scale," *Arch. Gen. Psychiatry*, vol. 12, no. 1, pp. 63–70, 1965.

[69] S. Wen, X. Meng, J. Chen, and E. Al., "Comparative Study on the Application of PHQ-9 and SDS in Patients with Screening for Depression in Emergency Department Waiting for Hospital Admission," *Sichuan Med. J.*, vol. 38, no. 2, pp. 5–9, 2017.

[70] Y. Lecrubier *et al.*, "The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI," *Eur. psychiatry*, vol. 12, no. 5, pp. 224–231, 1997.

[71] M. Zimmerman, J. H. Martinez, D. Young, I. Chelminski, and K. Dalrymple, "Severity classification on the Hamilton depression rating scale," *J. Affect. Disord.*, vol. 150, no. 2, pp. 384–388, 2013.

[72] A. J. Rush *et al.*, "The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression," *Biol. Psychiatry*, vol. 54, no. 5, pp. 573–583, 2003.

[73] H. Ellgring, *Non-verbal communication in depression*. Cambridge University Press, 2007.

[74] P. Ekman, "Facial action coding system," 1977.

[75] A. Pampouchidou *et al.*, "Automatic Assessment of Depression Based on Visual Cues: A Systematic Review," *IEEE Trans. Affect. Comput.*, vol. XX, no. c, pp. 1–27, 2017, doi: 10.1109/TAFFC.2017.2724035.

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

[77] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[78] J. P. Pestian *et al.*, "A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial," *Suicide Life ‐ Threatening Behav.*, vol. 47, no. 1, pp. 112–121, 2017.

[79] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016, doi: 10.1109/TAFFC.2015.2457417.

[80] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[81] L. Zhang and J. Driscol, "Evaluating Acoustic and Linguistic Features of Detecting Depression Sub-Challenge Dataset," no. 1, pp. 47–53, 2019.

[82] X. Li, "XMNLP: A Lightweight Chinese Natural Language Processing Toolkit," *GitHub*, 2018. https://github.com/SeanLee97/xmnlp.

[83] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.

[84] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?," in *International Conference on Machine Learning*, 2021, pp. 813–824.

[85] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv Prepr. arXiv1808.01340*, 2018.

[86] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[87] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social

withdrawal in depression: Evidence from manual and automatic analyses," *Image Vis. Comput.*, vol. 32, no. 10, pp. 641–647, 2014, doi: 10.1016/j.imavis.2013.12.007.

[88] Y. Yang, C. Fairbairn, J. F. Cohn, and A. Member, "Detecting Depression Severity from Vocal Prosody," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 142–150, 2013.

[89] American Psychiatric Association, "DSM-5 Self-Rated Level 1 Cross-Cutting Symptom Measure-Adult," 2013, [Online]. Available: http://psychiatry.org.

[90] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," *13th Annu. Conf. Int. Speech Commun. Assoc. 2012, INTERSPEECH 2012*, vol. 2, pp. 1058–1061, 2012.

[91] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[92] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Comput.*, vol. 13, no. 3, pp. 637–649, 2001.

[93] F. Eibe, M. A. Hall, and I. H. Witten, "The WEKA workbench. Online appendix for data mining: practical machine learning tools and techniques," in *Morgan Kaufmann*, 2016.

[94] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," 2015, [Online]. Available: http://arxiv.org/abs/1508.01991.

[95] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, [Online]. Available: http://arxiv.org/abs/1810.04805.

[96] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014.

[97] Y. Jiang, W. Li, M. S. Hossain, M. Chen, and A. Alelaiwi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Inf. Fusion*, vol. 53, no. February 2019, pp. 209–221, 2020, doi: 10.1016/j.inffus.2019.06.019.

[98] S. Song, L. Shen, and M. Valstar, "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," *Proc. - 13th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2018*, pp. 158–165, 2018, doi: 10.1109/FG.2018.00032.

[99] Z. Du, W. Li, D. Huang, and Y. Wang, "Encoding visual behaviors with attentive temporal convolution for depression prediction," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–7.

[100] L. Lin, X. Chen, Y. Shen, and L. Zhang, "Towards automatic depression detection: A bilstm/1d cnn-based model," *Appl. Sci.*, vol. 10, no. 23, pp. 1–20, 2020, doi: 10.3390/app10238701.

[101] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—Hybrid architectures," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 239–253, 2018.

[102] M. Rohanian, J. Hough, and M. Purver, "Detecting Depression with Word-Level Multimodal Fusion.," in *INTERSPEECH*, 2019, pp. 1443–1447.

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

[104] C. Sobin and H. A. Sackeim, "Psychomotor symptoms of depression," *Am. J. Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997.

Bochao Zou is a lecturer at the School of Computer and Communication Engineering, University of Science and Technology Beijing, China. He received the Ph.D. degree from Beijing Institute of Technology in 2018. He was a joint training Ph.d. student in Visual Attention Lab at Harvard Medical School and Brigham & Women's Hospital, Boston, USA from Sep. 2015 to Sep. 2017. His primary research area is affective computing with special interests in computational assessments of psychotic disorders. He is a member of IEEE.

Jiali Han. is a graduate student at Beijing Anding Hospital, Capital Medical University. Her research focuses on depression/bipolar disorder.

Yingxue Wang is a senior engineer of the National Engineering Laboratory for Risk Perception and Prevention, Beijing, China. She received the Ph.D. degree from Beijing Institute of Technology in 2017. Her current research interests include machine learning, speech signal processing, and their applications in affective computing.

Rui Liu is a research fellow of the National Clinical Research Centre of Mental Disorders, Beijing Anding Hospital, Beijing, China. She the Ph.D. degree from Southeast University in 2018. Her research focuses on medical data analysis.

Shenghui Zhao is an associate professor at Beijing Institute of Technology, Beijing, China. He received the Ph.D. degree in Information and Signal Processing at Beijing Institute of Technology in 1999. His research interests include speech and digital signal processing.

Lei Feng is an associate chief doctor of the National Clinical Research Centre of Mental Disorders, Beijing Anding Hospital, China. His research focuses on the early diagnosis and prediction of depression/bipolar disorder.

Xiangwen Lyu is a senior engineer and an expert of the National Engineering Laboratory for Risk Perception and Prevention. He received the Ph.D. degree in computer science at Nanjing University of Aeronautics and Astronautics in 2015. His current research interests include high-performance computing and affective computing.

Huimin Ma is a professor at the School of computer and communication, University of Science and Technology Beijing. She received the Ph.D. degree from Beijing Institute of Technology in 2001. She was an associate professor at Tsinghua University from 2006 to 2019. She is now the dean of the Department of Internet of Things and Electronic Engineering and the vice president of the Institute of artificial intelligence at USTB. She is also the secretary-general of China Society of Image and Graphics. Her research interests include 3D image cognition and simulation. In recent years, the researches were published in high-level journals (TPAMI, TIP, etc.) and international conferences (CVPR, NIPS, etc.).