Semi-structural inter
multimodal depress
automatic prelimi
depressive

Bochao Zou, *Member, IEEE*, Jiali Han, Yingx
Xiangwen Lyu,

Abstract—Depression is a common psychiatric disorder worldv
depression are not diagnosed, and most of them are not aw
automatic depression screening from behavioral indicators ha
multimodal depression corpus in Chinese since linguistic knowle
comprehensive survey with psychiatrists from a renowned psy
related to the diagnosis of depression. Then, a semi-structural in
undergone clinical diagnosis and professional assessment. Afte
analyzed between the two groups, statistically significant differen
of both single modal and multimodal fusion methods of depress
based fusion approach achieved the best performance. Finally
was made publicly available after de-identification and annot
research progress and practical applications of automatic depre

Index Terms—Affective computing; depressive disorder; multin

1 INTRODUCTION

DEPRESSION, otherwise known as major depressive disorder (MDD), is a common psychiatric disorder hat negatively impacts a person's way of thinking, feeling, nd behavior [1]. With the rapid development of society nd 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 milion, 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 lepression is about 6.9%, of which less than 10% of paients with depression have received professional assis-

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 Centr 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 Institut of Technology, Beijing, 100081, China. E-mail: shzhao@bit.edu.cn.

rview-based Chinese sion corpus towards nary screening of e disorders

xue Wang, Rui Liu, Shenghui Zhao, Lei Feng, and Huimin Ma*

wide. However, in China, a considerable number of patients with ware of their depression. Despite increasing efforts, the goal of s not been achieved. A major limitation is the lack of available edge is crucial in clinical practice. Therefore, we first carried out a chiatric hospital to identify key interview topics which are highly nterview study was conducted over a year with subjects who have er that, Visual, acoustic, and textual features were extracted and nces were observed in all three modalities. Benchmark evaluations sion assessment were also performed. A multimodal transformer-, the proposed Chinese Multimodal Depression Corpus (CMDC) ation. Hopefully, the release of this corpus would promote the ession screening.

nodal corpus; semi-structural interview

tance and treatment, and a considerable number of p tients are not aware of their depression [4]. On the oth hand, among the few patients who seek treatment in tim the first hospital most of them visit is not psychiatric ho pitals nor the psychiatric department of general hospita which is easy to cause misdiagnosis, and ultimately del the treatment. Consequently, the National Health Co mission of China issued the first action plan for the preve tion and control of depression, entitled "Action Plan f Explorations of Specialized Services for the Prevention a Treatment of Depression", on September 11, 2020, whi includes the routine screening of depression througho the country [5].

Screening and prevention of depression are of great si nificance. However, traditional questionnaire-bas screening of depression is facing problems of lacking we trained healthcare personnel since clinical interviewe based screening is labor-intensive and self-evaluati questions lack accuracy [6]. Many symptoms of depressi are considered observable [7]–[9]. The Diagnostic and St tistical Manual of Mental Disorders (DSM) is the standa of psychiatric diagnosis, which describes a series of audi visual behavioral indicators of depression [10]. Howev these indicators are often not fully considered wh screening, diagnosing, and evaluating depression [11]. T assessment of depression relies almost entirely on patien orally reported symptoms described in particular que

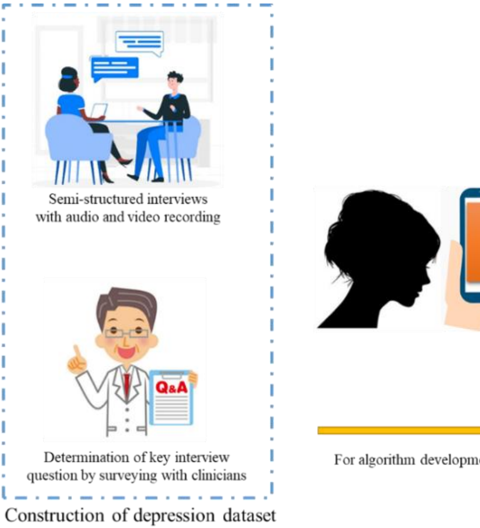


Fig. 1. Blueprint for automatic pre-screening of depressive dis

onnaires [12], such as the clinician-administered Hamil-n Rating Scale for Depression (HAMD) [13] and the self-port Patient Health Questionnaire (PHQ-9) [14]. Alt-ough these tools are useful, they neither include visual, oustic, or textual indicators of depression. To overcome is limitation, recent advancements in techniques for au-matic analysis of human behaviors, such as computer vi-on, speech signal processing, natural language under-anding, and multimodal learning could play an im-on, 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 per-

sorders through mobile application with build-in AI algorithms.

formed a comprehensive survey of MDD majored clini cians from a renowned psychiatric hospital in China to identify key interview questions which are highly relatec to the diagnosis of depression and can be used for AI-ena bled automatic depression screening, as shown in Fig. 1.

- We conducted semi-structural interviews based or key interview questions with participants of MDD anc Health Control (HC) subjects who have undergone clinica diagnosis and professional assessment of symptom sever

Health Control (HC) subjects who have undergone clinica diagnosis and professional assessment of symptom sever ity. During the interviews, the video and audio of partici pants were recorded simultaneously, and the audio wa transcribed into text through automatic transcription tool and proofreading.

• Significant features of visual, acoustic, and textua modalities between MDD and HC groups are reveale with statistical analysis. These significant difference among features confirm the feasibility of automatic de pression analysis with machine learning methods.

• Comprehensive benchmark evaluations are con ducted on the proposed dataset to provide a basis for an follow-up depression assessment research. A multimoda transformer-based fusion approach is applied in the do main of depression analysis which achieved the best resul during evaluation.

• The proposed Chinese multimodal dataset was mad publicly available after annotation and de-identification[1] To distribute the dataset publicly, we extract the feature of key questions related videos and audios with open source toolkits to eliminate personal information.

We believe that the release of this dataset will promot the research progress and practical applications of depres sion screening and assessment with core affective compu ting technologies, which could have significant scientifi value and broad application for societal questions of men tal health.

This paper is organized as follows. Section 2 reviews re ated 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 de-nd 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 n important role in depression detection, indicating the mportance of semantic information [41]–[43]. These explorations facilitate the interpretation of depression behavior ince it is obtained through multimodal behavior analysis of clinically matched control depression dataset. Thus, it is mportant to identify the most meaningful patterns of depressive behavior since behavior is associated with depresion-related symptoms in psychiatry studies [27].

2.2 Depression Assessment with Behavioral Indicators

There is an increasing number of studies on behavior indiators-based depression detection, and the research trend xtends 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 andmarks are common features for depression-related nalysis and can be combined with machine/deep learnng tools for depression detection. Pampouchidou et al. [34] gave a systematic review of visual cues based methods. Reent 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 feaure 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 mall sample number, such as in [19], they used pre-each speech segment, and then LSTM network with se attention mechanism was integrated to train the dow stream classification task. A latest review of deep learnir based depression analysis with audiovisual cues can found in [52]. Semantic information is also of great ir portance in depression detection, previous studies ha suggested superior performance of textual features [5 Contextual sentence embeddings, such as BERT [54], we shown to achieve better performance than that of wor level [41]. However, a recent study which used graph ne ral network to form the embedding of specific nodes word vectors showed that their method outperforms pr vious state-of-art methods by a substantial margin [46].

The results of several years' Audio/Visual Emoti Challenge and Workshop (AVEC) showed that metho based on multimodal fusion usually achieved better r sults [49], [55], [56]. In terms of fusion paradigm, early f sion, feature-level fusion, and late fusion were all explor by relevant research [47]. With the success of transforme in natural language understanding and computer visi tasks, transformer-based fusion methods have also be proposed as fusion methods and demonstrated their a vantages with temporal data by automatically alignir and capturing complementary features [57]. In summar a wide range of studies have been conducted in the field depression detection based on multimodal features, ar considerable progress has been made in performanc which provides a guarantee for the practicability of prelir inary screening for depression based on semi-structur

which provides a guarantee for the practicability of prelim-
inary screening for depression based on semi-structure
interviews.

2.3 Multimodal Datasets for Depression Assessment

Well-labeled multimodal recordings of clinically releva
behavioral differences between depressive and health
subjects are necessary for an automatic screening system
train classifiers [26]. Clinical datasets are hard to constru
since the difficulty in participant recruiting, and are us
ally public unavailable due to the confidentiality of patie
data. Table I shows a summary of interview-based datase
for depression assessment. We put these together to co
duct a thorough analysis to highlight the strength of t
proposed dataset.

Among them, the Distress Analysis Interview Corp
(DAIC-WOZ) [58], University of Pittsburgh depression d
taset (Pitt) [7], and Black Dog Institute depression datas
[28] are the three influential depression datasets. Speci
cally, the BlackDog dataset was collected in a depressio
specialized clinical research facility. Their interviews we
conducted with specific open-ended questions, such
portrayal of occasions in their life that had excited critic
feelings. Until now, the dataset has not been made pub
yet. The Pitt dataset contained 49 participants in a clinic
trial for the treatment of depression [7]. All their partic
pants met DSM-IV criteria for MDD. The severity of d
pression was evaluated with HAMD-17. This dataset
distributed upon request. The DAIC-WOZ dataset co
tains audio and facial features of depressed patients an
control subjects. The expert evaluated HAMD-17 and se
report PHQ-8 scores are provided for each patient. This a

TAB
SUMMARY OF INTERVIEW-BASED DAT

Dataset	Participant source	Subject No. (MDD/HC)	Selection Criteria	Clinical Diagnosis
DAIC-WOZ [58]	depression, PTSD, and anxiety	189(-/-)	-	No
Black Dog Institute [28]	Melancholia or MDD	60(30/30)	Clinical assessment	Yes (DSM)
Pittsburgh [7]	MDD	49(49/-)	HAMD>15	Yes (DSM)
Lin et al. [60]	depression and anxiety	35(18/17 for high & low distress)	PHQ-8⩾ 6.63	No
Jiang et al. [61]	Depression	12(12/-)	a 50% decrease in HAMD score	Yes
Guo et al. [62]	Depression	208(104/104)	PHQ-9⩾5	Yes
MODMA [63]	MDD	52(23/29)	PHQ-9⩾5	Yes (DSM)
Shen et al. [64]	Depression	162(30/132)	SDS⩾53	No
Proposed	**MDD**	**78(26/52)**	**HAMD-17 >17 or PHQ-9⩾9**	**Yes (DSM)**

PHQ-9≥9

here research assistants or a computer agent asked a se-
es of questions designed to identify depressive symp-
ms [58]. There are also depression detection datasets in
e AVEC [59], which is a series of competitions that have
en held for several years. In AVEC 2013, 2014, 2016, 2017,
d 2019, there were sub-challenges in depression detec-
on. The datasets in AVEC 2013 and 2014 were task-driven
havior observations of depressive people, which were
t interview-based. For 2016, 2017, and 2019, they all em-
oyed the subset of DAIC-WOZ dataset.

Besides the above three datasets, Lin et al. [60] intro-
uced a new audio-visual dataset containing full body vid-
s for distress detection. Currently, only a few studies at-
mpted to include the body modality which is worth ex-
oring. In terms of dataset construction, their participants
ere recruited online without clinical diagnosis. Both de-
ession and anxiety participants were included. The
morbidity of depression and anxiety makes it a big chal-
nge to distinguish between these two. For our purpose,
develop prescreening of MDD, including patients with
her psychotic disorders may introduce confounding fac-
rs. Visual and audio were recorded in their dataset but
thout text. Jiang et al.'s study [61] has a different re-
arch question with a cohort of 12 depressed patients
med at assessing the recovery, as well as the response to
ep brain stimulation treatment [65]. Their interview was
structured and data are not available. There are also
ore early studies that are not included in Table 1 since
ey are not available anymore, such as ORYGEN [66] and

BLE 1
TASETS FOR DEPRESSION ASSESSMENT

Interview Questions	Labels	Modality	Availability to Third Parties	Language
Open ended question	PHQ-8	A, V, T	Yes	Engli
Open ended question	QIDS-SR	A, V	No	Engli
From HAMD	HAMD	A, V	Upon request	Engli
Peer-support interview questions	PHQ-8 and GAD-7	V, A	Upon request	Engli
Unstructured	HAMD	V	No	Engli
Based on depression scales	PHQ-9	A, V	No	Chine
Based on depression scales	PHQ-9	A, EEG	Yes	Chine
Randomly selected questions	SDS	A, T	Yes	Chine
Semi structural questions identified with expert survey	**HAMD & PHQ-9**	**A, V, T**	**Yes (De-identified data and features)**	**Chine**

expert survey	PHQ-9	features)

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 question 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 difference [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] 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.

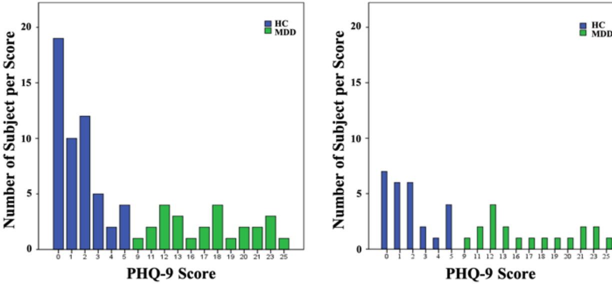
Therefore, we conduct a comparably large-scale Chinese multimodal depression study with semi-structural interviews under clinically validated diagnosis. As shown in Table I, the strength of the proposed dataset are highlighted in the following aspects: first, rigorous inclusion/exclusion criterion: clinical 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

Anding Hospital and the HCs were recruited from Beijing nstitute of Technology. In total, Seventy-eight subjects participated this experiment, which included 26 MDD paients (8 males, 18 females) with a mean age of 24.1 (SD = .04, age range 19-30 yr), and 52 HC (17 males, 35 females) vith 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 ubject. Subjects were asked before the experiment if they vould like to be video recorded, and the recorded video nay be published in research papers without any personal nformation in the consent form. Among all subjects, as hown in Table 2, 45 subjects have consented to video reording (19 MDD (14 females, mean age=23.6, SD=3.45) nd 26 HC (20 females, mean age=30.5, SD=11.92)), the rest vere audio-recorded only. The distribution of PHQ-9 cores of 78 subjects, as well as 45 subjects with video reording, are shown in Fig. 2. The Mini International Neuopsychiatric Interview (MINI) was employed to obtain he diagnosis [70]. All MDD participants met DSM criteria or major depression as assessed by professional psychiarists. Both HAMD-17 and PHQ-9 were assessed to serve s ground-truth labels for the development of automatic AI tools. PHQ-9 is a self-assessment questionnaire used to core nine DSM-IV criteria for depression. It is widely used s a tool for self-screening of depression. HAMD is a cliniian-rated tool composed of 17 items to quantify the severty of depression. The HAMD assessed the severity by exploring mood, guilt feeling, suicidal ideation, insomnia, nxiety, and somatic symptoms. Interviewers were experts



(a) (b)

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

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

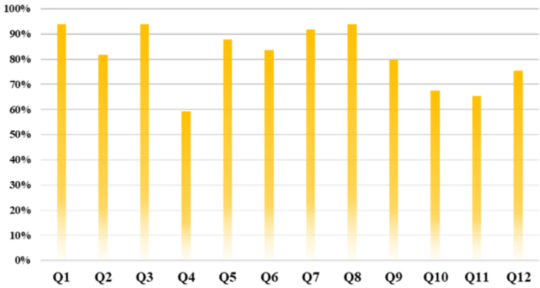
studies may have different criteria for depression and co trol grouping. For instance, studies of [62] and [6 adopted PHQ-9⩾5 for the depressed subject while Lin al. [60] used PHQ-9⩾6.63. Lin et al. grouped their partic pants according to the public norm. A score between 5 suggests mild depression which may require only watc ful waiting. Scores between 10-14 suggest moderate d pression severity that patients should have a treatme plan [14]. Instead of HAMD > 15 for moderate to seve depression used in Pittsburgh [7] dataset. We deploy t following criterion recommended by [71]: A HAMD sco of 18 or higher is generally considered to be moderate severe depression; a score of 8-17 indicates mild depre sion; a score of 7 or less indicates remission. These cuto scores are established with a large study of psychiatric ou patients with major depressive disorder and are current adopted in the clinical scenario in China. A PHQ-9 of 9 higher is semantically comparable to a HAMD score of or higher. As suggested by clinical psychiatrists, the scores are adopted as the selection criterion in our study

Only native Chinese-speaking participants were r cruited to reduce the differences caused by different la guage acquisition. Due to the limited availability of MD subjects, the HC group has more number of subjects and

TAB
KEY INTER

No.	
1	How are your appetite and weight change in the las
2	How's your health recently?
3	Sleep-related questions*
4	How often do you communicate with your friends r
5	How about your memory recently? Do you often fo
6	How are you interested in your current study or job
7	When do you feel tired or lack of energy? Does this
8	Have you ever thought of committing suicide or hu
9	Share your recent experience of feeling down, depr
10	What are the problems or worries in your life? And
11	When do you feel like you are a terrible failure, or l
12	When do you feel slower in your actions, thinking,

Note: Sleep-related problems may be asked differently depending on contexts
Will you wake up at night? Did you wake up too early? How long do you sl



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

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

d age biases are minimized to increase the statistical ower. Once subjects met the inclusion criteria, they were structed to proceed with the interview session.

Inclusion criteria of the MDD subjects were: (i) outpaent, age 18~65 years old, either sex; (ii) the diagnosis of DD in accordance with DSM-IV as (a) established by the sessing psychiatrist, and (b) confirmed with M.I.N.I.).0; (iii) a HAMD-17 score >17 or a PHQ-9 score ≥ 9; and) the informed consent was signed by the patient.

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

2 Experiment Setup and Apparatus

ch interview was conducted by one of two research asstants (RAs) as interviewers. During the interview, only terviewers and interviewees were in the room (Fig. 3). deo and audio streams were captured separately. Based the research literature on depression, we expect that the terpersonal nature of clinical interviews would improve e distinguishability across modalities [7]. A portable

BLE 3
VIEW TOPICS

Questions
st two weeks?
recently? What is your best friend's evaluation of you?
rget things?
? How is your concentration in your daily life?
s happen frequently?
rting yourself in any way? If so, what caused it?
ressed, or even desperate.
how do you deal with them?
have let yourself or your family down?
or speaking?

s, such as How is your sleep recently? How long does it take to fall asleep? leep every day? How easy it is to get a good quality of sleep?

(SONY HDR-CX680) for video recording. The camera i placed on a tripod behind the interviewer (in front of th interviewees). For each participant, the position of th camera was adjusted to ensure the optimal recording of fa cial areas. The audio recorder was placed on the desk at distance of approximately 60cm. During the semi-struc tural interview, the interviewer and interviewee sat face t face. The audio was low pass filtered at 75 kHz, and th frame rate of video recording was 50 fps. All subjects wer recorded using the same software and hardware apparatu Though MDD and HC subjects were recorded at differen sites, the room setting is the same (both are electromag

Though MDD and HC subjects were recorded at differen sites, the room setting is the same (both are electromag netic shielding rooms). Mobile phones were turned off o set to flight mode before recording. Participants wer asked to avoid touching the recording stick and table dur ing recording; prior to the key questions, the experimente recorded for about 3 minutes before the official recordin (did not involve the questions provided). All sessions wer recorded during office hours. Due to the difficulty of MDI subject recruitment, the data were collected over a year.

3.3 Data Collection

The interview was semi-structured, starting from neutra issues, aiming to establish a harmonious relationship an make participants feel comfortable, then proceeded to ke interview topics. The determination of interview topics i a key issue for the construction of a semi-structural inter view based depression corpus. The simple way is to follov the established questionnaires such as HAMD and PHQ-9 However, this is quite straightforward since psychiatrist usually do not strictly follow any questionnaire durin clinical practice. Considering the aim of this study, to buil an automatic pre-screening tool through digital interview with MDD patients that mimic the procedures of real clini cians, after a thorough discussion with senior clinicians we developed a list of key interview questions based o the topics that clinicians may talk about during their prac tice, as well as questionnaires (PHQ-9, QIDS-SR [72], an HAMD). This process ended up with 35 questions by prob ing appetite, sleep, physical exercise, health condition job/study, social interaction, memory, concentration, sui

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 land-

Head Poses	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.

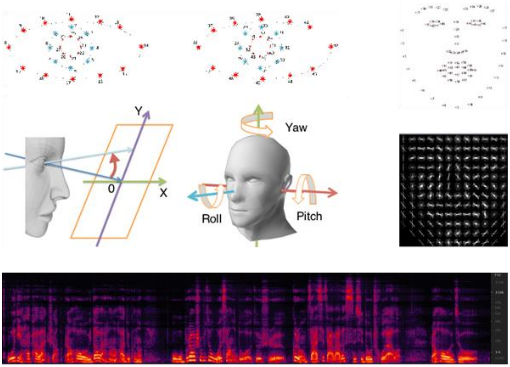
consume a large amount of time, and therefore not suitable for preliminary screening.

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

(a)

(b)



(c) 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 visual features (2D&3D facial landmarks, eye landmarks, gaze dire tions, histogram of gradients, head rotations), (b) audio recording, v alized as spectrogram, and (c) transcription of the patient A s spee (both in Chinese and English).

to stimulate spontaneous speech, facial expressions, ar related body gestures. The semi-structured interview w conducted by one of the trained RAs, and they were awa of the depression status of the subjects.

4 Annotation & Feature Extraction

4 ANNOTATION & FEATURE EXTRACTION

4.1 Dialogue Segmentation

The audio and video recording of interviews included k topics about events and symptoms related to depressic as well as neutral questions designed to build rapport i teractions. Only key interview questions were located ar cropped from the original videos and audios with questic information to support the development of AI tools for a tomatic screening. More specifically, we located the sta and end of each key question by reviewing the whole a dio stream and synchronizing each video clip simultar ously.

4.2 Interview Transcription

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

4.3 De-identification

All transcribed interviews were annotated to remove ide tifying information. Utterances that mention person names, specific addresses, workplaces, and can be used narrow the scope of the event were tagged and eliminate ata annotated in the corpus does not contain protected alth information. The facial signals are features, such as cial landmarks, eye landmarks, gaze angles, head movements, HOG features, AU intensity, and AU occurrence, hich do not contain enough information for personal entification.

4 Visual Feature Extraction

DD can be manifested through a variety of visual signs 3]. Such as changes in the activities of facial muscles, as ell as eye gaze direction, often imply the persistent negave thoughts and feelings of sadness that characterize deession. Action Units (AUs) were proposed and defined Ekman et al. [74] which describe the coordinated activies of facial muscle groups corresponding to specific faal expressions. AUs can be used to describe the changing aracteristics of facial expression in depression since DD patients often have poor expression ability. In addion, the occurrence of specific facial movements (smile, rner of mouth down, etc.) described by specific AUs (e.g. U12) is directly related to depression. Thus, various studs have applied AUs to the automatic assessment of deession and achieved promising results [75]. Gaze angle, ad pose, and facial landmarks have also been applied to epression assessment. The average gaze angle and ange of gaze direction of the eye were also used as the etection characteristics of depression [76]. Facial landarks can be applied in facial expression analysis. Therere, AUs, eye gaze, head pose, and facial landmarks were

rks can be applied in facial expression analysis. There-
re, AUs, eye gaze, head pose, and facial landmarks were
tracted as visual features in the proposed depression
rpus.

As mentioned in Section 3.1, 19 MDD and 26 HC have
nsented to video recording among all subjects. Auto-
atic annotation of non-verbal features was carried out us-
g an open-source facial behavior analysis toolkit called
penFace (version 2.2.0) [77]. OpenFace is adopted since it
the state-of-art open-sourced tool for facial movement
alysis, and is widely used in depression analysis, which
ay potentially facilitate corpus usage [34], [60]. The de-
ils of visual feature sets are shown in Table 5. The Open-
ce features were frame-level but were averaged into sen-
nce level with partitioned interview parts for subsequent
atistical analysis and evaluation.

5 Acoustic Feature Extraction

peech conveys non-verbal information on the depressive
ate of the speaker and the corresponding acoustic fea-
res are affected by depression [38]. For example, relevant
udies have observed certain vocal changes in MDD pa-
ents, such as increased pause time and impaired fluency
0]. People with mental disorders also have shown dis-
rbances in prosody [36]. Recently, machine learning with
oustic features has achieved a high-precision classifica-
on of suicidal thoughts in MDD patients [78]. The ex-
nded Geneva Minimalistic Acoustic Parameter Set (eGe-
APS) [79] is a quite standard feature set that is usually
opted as acoustic features in depression detection chal-
nges such as AVEC [59]. Therefore, eGeMAPS were
opted as the acoustic feature set in the proposed corpus.

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.

tagging as well as sentiment analysis were achieved with Xmnlp (version 0.3.1) [82] which is an open-source light weight 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 ex

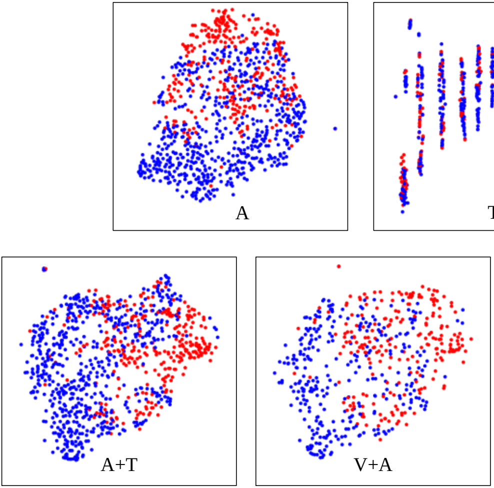


Fig. 6. T-SNE 2D visualization, each dot represents one partitioned i
from data of all 78 participants. As to V and V related feature

he model URL is given in [3]. Timesformer is a state-of-art nodel 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 nodeling. Furthermore, for deep representation of audio, 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

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.

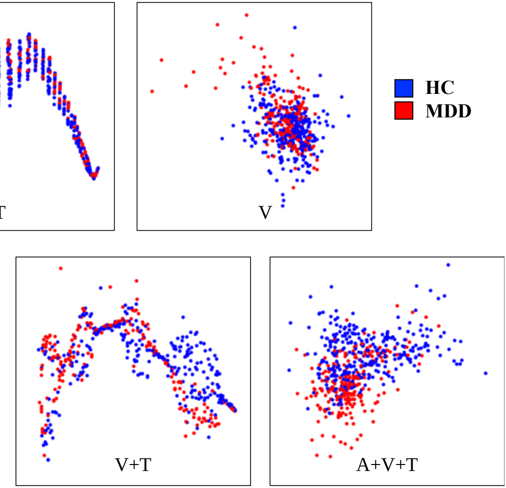
5 Statistical Analysis & Benchmark Evaluation

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



nterview section (For modality of A, T, and A+T, results were obtained
s, results were from the 45 participants with video recordings).

and vertical gaze angles, SD represents an estimate of t
variance of gaze angle), mean intensity of 17 AUs, the o
currence rate of 18 AUs, see [77] for a detailed explanati
of features. A three-way MANCOVA analysis was co
ducted to test for the main effects of subject group, inte
view question, and gender. The results reveal the main e
fects of group (Wilks' Lambda $F(49, 436) = 18.14$, p<.00
$\eta_p^2 = 0.67$) and gender (Wilks' Lambda $F(49, 436) = 9.54$
p<.001, η_p^2=0.52) as well as an interaction effect of grou
and gender (Wilks' Lambda $F(49, 436) = 9.72$, p<.00

p<.001, η_p^2=0.52) as well as an interaction effect of grou and gender (Wilks' Lambda $F(49, 436) = 9.72$, p<.00 η_p^2=0.52). This indicates salient visual differences betwe groups and genders and the significant visual features MDD vary with gender. For individual visual feature however, none of them has a large effect size. Only occu rence rate of AU06_c (Cheek Raiser, $F(1, 530)$=14.36, p<.0 η_p^2=.098) has a significant difference with a medium effe The average intensity of AU01 (Inner Brow Raiser), AU (Outer Brow Raiser), AU04 (Brow Lowerer), AU12 (L Corner Puller), AU20 (Lip stretcher), AU23 (Lip Tightene and AU25 (Lips part), as well as the occurrence rate AU04 (Brow Lowerer), AU12 (Lip Corner Puller), AU (Dimpler), AU17 (Chin Raiser), AU20 (Lip stretcher), AU (Lip Tightener), AU25 (Lips part), AU28 (Lip Suck) we found differ significantly between MDD and HC grou with a small effect size. Although gaze-related paramete do not show a large effect size of significant difference, ve tical gaze angle has a p-value less than 0.001, which d scribes the more downward angle of gaze for MDD.

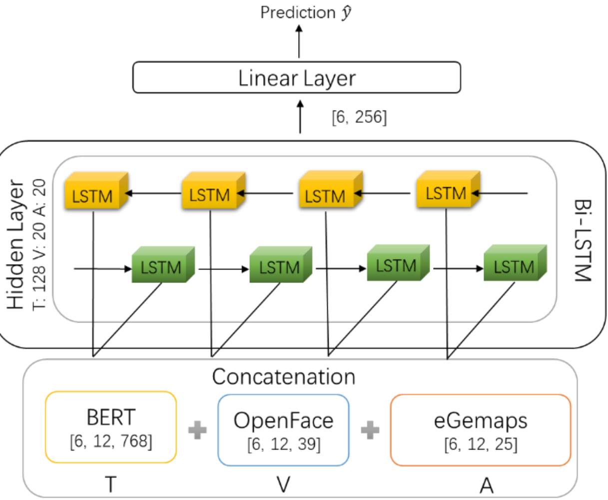
These results are consistent with previous findings [23], [87], which found reduced affiliative expressions a increased non-affiliative expressions in MDD. In particul both studies found the higher average intensity of AU0 AU04 which is associated with sadness emotion, and low intensity of AU06 which is a common expression for ha piness [75]. All significant AUs are around the eye a mouth areas. The detailed statistics of visual features b tween MDD and HC are given in Table S1 of the suppl mentary materials.

1.2 Acoustic feature analysis

or acoustic features, a three-way MANCOVA analysis reals 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, 5) =29.78, p<.001, η_p^2=0.47) as well as an interaction efct of group and gender (Wilks' Lambda F(25, 845) = 14.30, .001, η_p^2=0.30). This indicates salient acoustic parameters fferences between groups and genders, and the significant acoustic features of MDD vary with gender. For individual acoustic parameters, only ones with large effect zes were reported as significant features. Although main fects of subject group were found on most acoustic parameters (except mfcc1, Harmonic difference H1-H2, H1-3, and Alpha Ratio), only mfcc4 (F(1, 915)=245.22, p<.001, =0.21), F1_Bandwidth (F(1, 915)=272.79, p<.001, =0.23), F2_Bandwidth (F(1, 915)=333.86, p<.001, =0.27), Hammarberg Index (F(1, 915)=430.63, p<.001, =.32) Spectral Slope 0-500 Hz (F(1, 915)=920.10, p<.001, =.50) were significantly different between two groups IDD and HC) with large effect sizes. To further assess the fects of question and gender on the five significant features, the MANCOVA results of the five features were anyzed. No main effect of question and gender, as well as e interaction effect between question and group, gender d group, and question, gender, and group was found (all s<0.14).

Loudness represents the estimate of perceived signal innsity and has been studied intensively in the detection of epression[81], [88]. The results of this study show that

nsity and has been studied intensively in the detection of
epression[81], [88]. The results of this study show that
tch and loudness are both lower in MDD group, which
e in agreement with previous studies [35] and clinical ob-
rvation [89] that MDD subjects are generally believed to
ve a lower sound volume than HC subject. MFCC is Mel-
equency Cepstral Coefficients that represent acoustic
act changes [35]. Hammarberg Index is the ratio of the
ghest energy peak of 0-2 kHz region to that of 2–5 kHz
gion. The spectral slope is the linear regression slope of
e logarithmic power spectrum within two given fre-
uency bands [79]. Both Hammarberg Index and spectral



slopes are previously reported to have a significant corre lation with depression severity [90]. Significant difference in formant 1, 2 bandwidths were also found between th two groups. Formant features have already been shown t capture information useful in distinguishing between th two classes [39] and have also been linked with emotiona and cognitive information. The detailed statistics of 2 acoustic features between MDD and HC groups are show in Table S2 of the supplementary materials.

5.1.3 Textual feature analysis

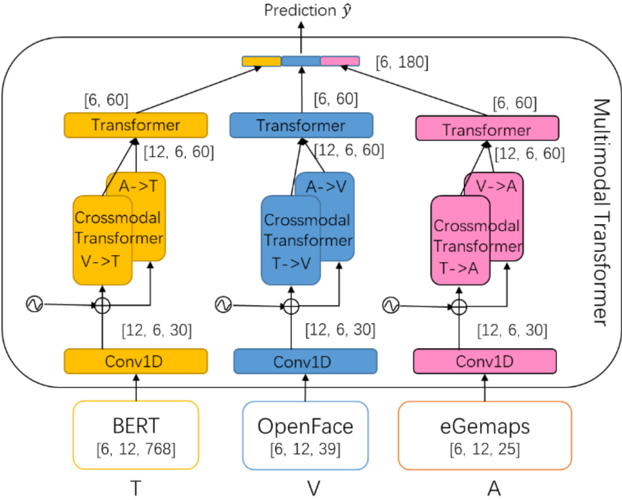
As to textual features, a three-way MANCOVA analysi reveals main effect of subject group (Wilks' Lambda F(10 863)= 14.90， p=.001, η_p^2=0.018) and interview questio (Wilks' Lambda F(110, 6471.45) = 1.46，p<.001, η_p^2=0.147 no significant effects of gender and interactions amon them (p=.11). MDD subjects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

5.2 Benchmark Evaluation

Benchmark evaluations of the proposed dataset were als conducted to provide a basis for any follow-up multimoda depression assessment research. The clinical diagnosis o MDD was considered as the ground truth for the classifi cation task and the scores of PHQ-9 questionnaire were th ground truth for the regression task. Previously extracte cation 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



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

Furthermore, we also evaluated two more deep learn-ng based methods on the proposed dataset for both de-pression classification and regression as shown in Fig.7 nd Fig. 8. The first one is an early fusion bi-directional LSTM (EF-Bi-LSTM) [94] model which takes the concate-nation of textual, visual, and acoustic features as input. The visual and acoustic features were extracted with Openface nd OpenSmile, respectively, same as other baseline meth-ods. But the textual feature is encoded with pretrained Chi-nese BERT [95] embedding. Instead of early fusion, the sec-ond one is based on more advanced multimodal trans-ormer fusion [57] which can handle inherent data non-lignment and long-range dependencies across modalities. The layers and parameters settings are illustrated in the wo figures. For reproducibility, we share the code of Bi-LSTM and MuIT methods on the GitHub website: ttps://github.com/CMDC-corpus/CMDC-Baseline. For both models, we used Adam [96] to optimize with a batch ize of 6, and a learning rate of $1e^{-4}$. To overcome overfit-ing, dropout layers were inserted in Bi-LSTM and trans-

ize of 6, and a learning rate of 1e. To overcome overfit-
ing, dropout layers were inserted in Bi-LSTM and trans-
ormer layers with a dropout rate of 0.1. The number of

TAB

BENCHMARK EVALUATION OF DEPRESSION

	Metric	SVM (Linear)	SVM (SMO)
A	Precision	0.78	*0.92*
	Recall	0.78	**0.91**
	F1	0.78	**0.91**
	AUROC	0.77	*0.90*
T	Precision	0.71	0.84
	Recall	0.71	0.84
	F1	0.71	0.84
	AUROC	0.71	*0.84*
V	Precision	0.61	0.69
	Recall	0.60	0.69
	F1	0.60	0.69
	AUROC	0.60	0.67
A+T	Precision	0.90	*0.92*
	Recall	*0.91*	*0.91*
	F1	*0.91*	*0.91*
	AUROC	*0.90*	*0.90*
V+A	Precision	0.78	**0.89**
	Recall	0.78	**0.87**
	F1	0.78	*0.86*
	AUROC	0.77	*0.84*
V+T	Precision	0.72	*0.87*
	Recall	0.71	*0.87*
	F1	0.71	*0.87*
	AUROC	0.71	*0.86*
A+V+T	Precision	*0.91*	*0.91*
	Recall	**0.91**	*0.89*
	F1	**0.91**	0.89
	AUROC	**0.90**	0.87

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

hidden units for A, V, T are 20, 20, 128 respectively and t
number of training epochs was set as 200. The paramete
were chosen accordingly for different fusion methods (r
fer to the shared code for details). As shown in the last tw
columns of Table 6, the (EF) Bi-LSTM model perforn
comparably to other baseline methods. But the multimod
transformer fusion model generally performs better th
other fusion methods. As shown in the above results, t
depression classification performance almost reaches t
bottleneck with an F1 score of 0.95. Therefore, We furth
conducted a more difficult regression task.

5.2.2 Regression task

For the regression task, mean absolute error (MAE), ro
mean squared error (RMSE), and Pearson correlation coe
ficients were reported as evaluation metrics. The mea
PHQ-9 score computed on the training set for predicti
was conducted as a trivial baseline. A MAE of 7.03 and
RMSE of 7.84 are the trivial baseline results for 45 subjec
with all modalities,. Linear regression, Support vector r
gression, Multi layer perceptron, and K nearest neighbo
hood were also adopted as the baseline regressors wi
stratified 5-fold cross-validation, all were implement
with default parameters in Weka. As listed in Table
Some of the machine learning methods are even wor
than the trivial baseline, but the two deep learning mode
always outperform it. The best result is achieved by mul
modal transformer fusion of acoustic and textual moda
ties with MAE 3.66, RMSE 4.59, and Pearson 0.72. The f

modal transformer fusion of acoustic and textual modali-
ties with MAE 3.66, RMSE 4.59, and Pearson 0.72. The f

BLE 6
CLASSIFICATION WITH VARIOUS MODALITIES

Logistic	Naïve Bayes	Bi-LSTM (Early Fusion)	MulT
0.85	0.89	**1.00**	—
0.84	*0.89*	0.83	—
0.84	0.89	**0.91**	—
0.77	0.89	**0.92**	—
0.93	0.85	*0.87*	—
0.93	0.80	*0.90*	—
0.93	0.78	*0.88*	—
0.99	*0.84*	0.83	—
0.73	0.60	**1.00**	—
0.73	0.60	*0.71*	—
0.73	0.60	**0.83**	—
0.70	0.57	**0.86**	—
0.92	0.91	**0.97**	0.87
0.91	0.89	*0.91*	**0.96**
0.91	0.89	**0.94**	*0.91*
0.89	0.86	**0.91**	0.87
0.80	0.82	0.83	*0.87*
0.80	0.82	*0.83*	**0.87**
0.80	0.82	0.83	**0.87**
0.74	0.80	0.83	**0.87**
0.80	0.66	0.82	**1.00**
0.80	0.67	**0.89**	0.83
0.80	0.66	0.85	**0.91**
0.78	0.69	0.82	**0.91**
0.82	0.84	0.87	**0.97**
0.82	0.84	*0.89*	0.85
0.82	0.84	0.88	**0.91**
0.82	0.81	0.82	*0.88*

TABL
BENCHMARK EVALUATION OF DEPRESSION

	Metric	Linear Regression	SVR (SMOreg)
A	MAE	8.10	7.98
	RMSE	11.6	11.93
	Pearson	0.48	*0.53*
T	MAE	6.77	6.65
	RMSE	10.25	10.58
	Pearson	0.48	0.50
V	MAE	7.09	6.81
	RMSE	9.57	9.50
	Pearson	*0.44*	**0.47**
A+T	MAE	7.00	6.10
	RMSE	10.26	8.71
	Pearson	0.38	0.48
V+A	MAE	6.78	6.15
	RMSE	9.74	8.70
	Pearson	0.40	0.45
V+T	MAE	5.41	5.13
	RMSE	6.74	6.28
	Pearson	0.62	**0.67**
A+V+T	MAE	5.80	5.31
	RMSE	7.97	7.04
	Pearson	0.50	0.57

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	0.88	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.

on of three modalities also achieved relatively good re-
lts. Similar to the classification task, the fusion of multi-
odal generally perform better than the unimodal meth-
ls.

For further direction, more advanced representation
arning techniques for feature extraction, embedding, and
ultimodal fusion techniques are worth exploring for bet-
r performance [97].

2.3 Evaluation of baseline methods on DAIC

o further analyze the benchmark evaluation, we also eval-
ted the two baseline deep-learning methods on the pub-

LE 7
REGRESSION WITH VARIOUS MODALITIES

MLP	KNN	Bi-LSTM (Early Fusion)	MulT
6.26	*5.78*	**4.59**	—
8.83	*7.94*	**6.14**	—
0.48	0.46	**0.64**	—
4.77	5.96	**3.81**	—
6.47	8.08	**5.35**	—
0.64	0.59	*0.63*	—
6.73	9.13	**5.77**	—
9.02	11.79	**7.70**	—
0.33	-0.05	0.28	—
5.34	5.27	*3.68*	**3.66**
7.12	7.15	*4.62*	**4.59**
0.53	0.60	*0.71*	**0.72**
7.09	**5.04**	5.18	*5.08*
12.63	7.66	*6.43*	**6.02**
0.35	*0.57*	0.55	**0.60**
17.83	7.15	*4.76*	**4.61**
83.59	10.03	*5.75*	**5.56**
0.19	0.22	0.54	*0.64*
5.89	5.89	*4.55*	**4.32**
7.60	8.38	*5.67*	**5.61**
0.39	0.50	*0.68*	**0.72**

implemented two deep learning models for both classifica
tion and regression tasks with the same feature extractio
of V and T (with pretrained English Bert model) modalitie
For audio, the DAIC provides 74 audio features extracte
with the COVAREP toolbox (v. 1.3.2) [103] on every 10m
segment. The mean, max, and min of features on parti

with the COVAREP toolbox (v. 1.3.2) [103] on every 10m
segment. The mean, max, and min of features on parti
tioned question-level audio were calculated for tempora
aggregation. The training set and the validation set o
DAIC were combined for training, and the results were ob
tained on the test set. Since DAIC is widely used in the lit
erature, we also listed some previous results as a compari
son. As shown in Table 8, the performance of the two meth
ods are close to the previous state-of-art. By comparing Ta
ble 8 with Tables 6 and 7, we can observe that both the clas
sification and regression performance on the propose
CMDC are comparable but slightly better than that on th
DAIC, which may be due to the well-defined semi-struc
tural interview. Considering the classification accuracy o
the proposed corpus, the feasibility of preliminary screenin
based on multimodal behavior signals thus could somehow b
guaranteed.

6 Discussion & Conclusion

Previous studies found that depressed patients differ fron
the HC group in observable behavior modalities includin
visual, acoustic, and verbal signals [27], [34], [35], [38], [88]
Such as visual signals from the facial area show that psy
chological distress is predicted by larger downward angle
of gaze, shorter average duration, and less intensity o
smiles [33]. Depressed speech has found several distin
guishing acoustic features as indicators of disease severit
and treatment efficacy [104]. Therefore, a depression cor
pus with multidimensional behavior observations woul
promote the research of automatic depression detection
enabled by machine learning methods for the fusion o

tronger 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.

aset 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 with core affective computing technologies.

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

ACKNOWLEDGMENT

This work was supported by the National Natural Scien Foundation of China (No. U20B2062, U19B2032). The a thors would like to thank Guangyao Sun, Jingyu Hou, Xu song Wang, and Yubo An for their generous assistan during data collection and processing.

REFERENCES

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

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

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

[4] Y. Chen *et al.*, "Patterns and correlates of major depressi in Chinese adults: a cross-sectional study of 0.5 million m and women," *Psychol. Med.*, vol. 47, no. 5, pp. 958–970, 20

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

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

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

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

[8] S. Gao, V. D. Calhoun, and J. Sui, "Machine learning in ma
depression: From classification to treatment outco
prediction," *CNS Neurosci. Ther.*, vol. 24, no. 11, pp. 10
1052, 2018.

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

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

[11] A. F. Schatzberg, "Scientific issues relevant to improving t
diagnosis, risk assessment, and treatment of ma
depression," *Am. J. Psychiatry*, vol. 176, no. 5, pp. 342–3
2019, doi: 10.1176/appi.ajp.2019.19030273.

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

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

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

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

[16] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depressi
with audio/text sequence modeling of interviews," *Pr
Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, v

7] 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.

3] 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.

9] 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.

0] 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.

1] 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.

2] 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.

3] 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.

4] 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.

5] S. Gao and V. D. Calhoun, “Machine learning in major

10.1109/TAFFC.2017.2650899.

5] 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.

6] 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.

7] 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.

8] 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.

9] 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.

0] 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.

1] 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.

2] 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.

International Conference and Workshops on Automatic Face an Gesture Recognition (FG), 2013, pp. 1–8.

[34] A. Pampouchidou *et al.*, "Automatic Assessment o 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 depresse 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 depressio symptoms using an online cross - sectional data collectio 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 ris assessment using speech analysis," *Speech Commun.*, vol. 7 pp. 10–49, 2015, doi: 10.1016/j.specom.2015.03.004.

[39] D. J. France and R. G. Shiavi, "Acoustical properties o speech as indicators of depression and suicidal risk," *IEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, 2000, do 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.*, vo 207, pp. 359–366, 2017.

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

[42] S. Guohou, Z. Lina, and Z. Dongsong, "What reveals abou depression level? The role of multimodal features at the leve of interview questions," *Inf. Manag.*, vol. 57, no. 7, p. 10334

depression level? The role of multimodal features at the leve of interview questions," *Inf. Manag.*, vol. 57, no. 7, p. 10334 2020, doi: 10.1016/j.im.2020.103349.

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

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

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

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

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

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

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

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:

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.

approach to short-term detection of mood disorder," *Pr*
Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, v
08-12-Sept, no. October 2017, pp. 1452–1456, 2016, d
10.21437/Interspeech.2016-620.

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

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

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

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

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

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

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

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

[76] S. Scherer, G. Stratou, and L.-P. Morency, "Audiovisu
behavior descriptors for depression assessment,"
Proceedings of the 15th ACM on International conference

behavior descriptors for depression assessment," Proceedings of the 15th ACM on International conference multimodal interaction, 2013, pp. 135–140.

[77] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Moren "Openface 2.0: Facial behavior analysis toolkit," in *2018 1 IEEE International Conference on Automatic Face & Gest Recognition (FG 2018)*, 2018, pp. 59–66.

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

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

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

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

[82] X. Li, "XMNLP: A Lightweight Chinese Natural Langua Processing Toolkit," *GitHub*, 20 https://github.com/SeanLee97/xmnlp.

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

[84] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Ti Attention All You Need for Video Understanding?," *International Conference on Machine Learning*, 2021, pp. 81 824.

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

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

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.

3] 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.

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

0] 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.

1] 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.

2] 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.

3] 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.

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

5] 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.

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

7] Y. Jiang, W. Li, M. S. Hossain, M. Chen, and A. Alelaiwi, "A snapshot research and implementation of multimodal

7] 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.

3] 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.

9] 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.

00] 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.

01] 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.

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

03] 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.

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

chao Zou is a lecturer at the School of Computer and Communica-

2018. He was a joint training Ph.d. student in Visual Attention Lab a Harvard Medical School and Brigham & Women's Hospital, Boston USA from Sep. 2015 to Sep. 2017. His primary research area is affec tive computing with special interests in computational assessments o psychotic disorders. He is a member of IEEE.

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

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

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

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

Lei Feng is an associate chief doctor of the National Clinical Researc Centre of Mental Disorders, Beijing Anding Hospital, China. His re search focuses on the early diagnosis and prediction of depression/b polar disorder.

Xiangwen Lyu is a senior engineer and an expert of the National En gineering Laboratory for Risk Perception and Prevention. He receive the Ph.D. degree in computer science at Nanjing University of Aero nautics and Astronautics in 2015. His current research interests in clude high-performance computing and affective computing.

clude high-performance computing and affective computing.

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