Analysis of Automated Clinical Depression Diagnosis in a Chinese Corpus

Kaining Mao 1, Deborah Baofeng Wang 2, Tiansheng Zheng 2, Rongqi Jiao 2, Yanhui Zhu 2, Bin Wu 2, Lei Qian 2, Wei Lyu 2, Jie Chen 2, and Minjie Ye 2

1University of Alberta
2Affiliation not available

October 31, 2023

Abstract

Depression clinical interview corpora are essential for advancing automated depression diagnosis. While previous studies have used written speech material in controlled settings, these materials do not accurately represent spontaneous conversational speech. Additionally, self-reported measures of depression are subject to bias, making the data unreliable for training models for real-world scenarios. This study introduces a new corpus of depression clinical interviews collected directly from a psychiatric hospital, containing 113 recordings with 52 healthy and 61 depressive patients. The subjects were examined using the Montgomery-Asberg Depression Rating Scale (MADRS) in Chinese. Their final diagnosis was based on medical evaluations through a clinical interview conducted by a psychiatry specialist. All interviews were audio-recorded and transcribed verbatim, and annotated by experienced physicians. This dataset is a valuable resource for automated depression detection research and is expected to advance the field of psychology. Baseline models for detecting and predicting depression presence and level were built, and descriptive statistics of audio and text features were calculated. The decision-making process of the model was also investigated and illustrated. To the best of our knowledge, this is the first study to collect a depression clinical interview corpus in Chinese and train machine learning models to diagnose depression patients.
Analysis of Automated Clinical Depression Diagnosis in a Chinese Corpus

*Kaining Mao, **Deborah B. Wang, Tiansheng Zheng, Rongqi Jiao, Yanhui Zhu, Bin Wu, Lei Qian, Wei Lyu, †Jie Chen, Fellow, IEEE, and ‡Minjie Ye

Abstract—Depression clinical interview corpora are essential for advancing automated depression diagnosis. While previous studies have used written speech material in controlled settings, these materials do not accurately represent spontaneous conversational speech. Additionally, self-reported measures of depression are subject to bias, making the data unreliable for training models for real-world scenarios. This study introduces a new corpus of depression clinical interviews collected directly from a psychiatric hospital, containing 113 recordings with 52 healthy and 61 depressed patients. The subjects were examined using the Montgomery-Asberg Depression Rating Scale (MADRS) in Chinese. Their final diagnosis was based on medical evaluations through a clinical interview conducted by a psychiatry specialist. All interviews were audio-recorded and transcribed verbatim, and annotated by experienced physicians. This dataset is a valuable resource for automated depression detection research and is expected to advance the field of psychology. Baseline models for detecting and predicting depression presence and level were built, and descriptive statistics of audio and text features were calculated. The decision-making process of the model was also investigated and illustrated. To the best of our knowledge, this is the first study to collect a depression clinical interview corpus in Chinese and train machine learning models to diagnose depression patients.

Index Terms—Emotional corpora, Machine learning, Multimodal systems, Nonverbal signals, Sentiment analysis

I. INTRODUCTION

Major depressive disorder (MDD), also known as clinical depression, is a prevalent yet serious mental disorder that causes significant emotional distress and impacts daily functioning [1], [2]. MDD has become one of the most common and costly disorders worldwide in recent years, with a higher prevalence in low and middle-income populations with fewer economic and social resources [3], [4].

D. Wang, T. Zheng, L. Qian, W. Lyu and M. Ye are with The Affiliated Kangning Hospital of Wenzhou Medical University, Zhejiang Provincial Clinical Research Center for Mental Disorders, Ouijiang Laboratory (Zhejiang Lab for Regenerative Medicine, Vision and Brain Health), Wenzhou, Zhejiang 325000, China, Key Laboratory of Alzheimer’s Disease of Zhejiang Province, Institute of Aging, Wenzhou Medical University, Wenzhou, Zhejiang, China.

K. Mao and J. Chen are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, T6G 2R3.

R. Jiao, Y. Zhu, B. Wu are with Wenzhou Medical University, Wenzhou, Zhejiang 325000, China.

* These two authors contributed equally to this work.
† Co-corresponding authors

The World Health Organization (WHO) predicts that clinical depression will be the second most debilitating disease by 2030, following only cardiovascular diseases [5]. Early-stage depression screening and follow-up psychotherapy are crucial for individuals with high-risk mental health conditions [6]. However, delivering early-stage mental health interventions to vulnerable and stigmatized populations, particularly those in financial hardship or concerned about privacy, remains challenging.

In addition, depression diagnosis currently lacks a reliable “gold standard” method, which leads to inconsistencies in treatment depending on an individual clinicians’ experiences and intuition. Artificial Intelligence (AI) has the potential to revolutionize the diagnosis and prognosis of diseases, especially by making early-stage mental health screening a reality. The role of datasets cannot be overlooked in advancing AI models. Recent papers published in TBioCAS and IEEE CASS journals highlight the importance of datasets. For example, Zhang et al. conducted experiments on respiratory sound classification using their own SPRSound database [7]. Jiang et al. presented a Standardized Assessment of Underwater Image dataset and proposed a new metric for image quality assessment [8]. The availability of these datasets presents significant opportunities for researchers and practitioners in a variety of fields to develop innovative solutions to various problems.

Although significant progress has been made in automating depression diagnosis, previous studies have primarily used non-clinical datasets. These datasets are valuable for researchers as they provide ample training data and insights for those without the resources to collect and label their datasets. They also serve as a benchmark for performance evaluation, allowing researchers to compare their models with others. Previous studies have investigated music-induced [9], video-induced [10], [11] and mixed emotion induction methods [12]. However, there are still challenges to implementing and deploying depression detection systems in real-world applications. For instance, existing datasets ignore the critical aspect that emotions are typically context-based. To address this issue, there is a need for interactive multimodal datasets collected through interviews conducted in a clinical setting between patients and physicians. In this scenario, patient’s emotions rely on verbal and non-verbal communication with physicians. The primary objective of this study is to investigate the effectiveness of semantic and prosodic features in evaluat-
ing depression risk. To ensure the collection of a high-quality dataset, the data collection process must be controlled and standardized. Therefore, we conducted interviews in Chinese between clinicians and outpatients, and the patients were evaluated using the Montgomery-Asberg Depression Rating Scale (MADRS) [13]. The audio features, such as formant frequency $F_0$ and normalized amplitude quotient (NAQ), were then extracted using the collaborative voice analysis repository (COVAREP) toolbox, and the interview recordings were transcribed verbatim using an audio transcription application programming interface (API) developed by iFlyTek for subsequent analysis [14]. Research assistants majoring in psychology corrected errors in the transcripts, such as words with the same pronunciation but different meanings. The dataset included both the interview recordings and their transcripts. To the best of our knowledge, this is the first multimodal clinical distress interview corpus with over 100 subjects in Chinese.

In this paper, we address the challenges and limitations of existing datasets by introducing the Wenzhou Kangning dataset, an audio-text dataset of clinically annotated depression severity. Our study fills a gap in this area by addressing the scarcity of authentic depression interview datasets collected in clinical settings. The availability of our dataset not only facilitates the development of automated depression diagnosis models but also empowers researchers to explore novel approaches and techniques in the field. With mental health concerns on the rise globally, our study holds immense significance in promoting early detection, personalized care, and improved outcomes in mental healthcare. Our analysis shows a significant difference in the audio duration, and individual sentence word counts between healthy and depressive patients, indicating that linguistic cues can be an effective predictor of a subject’s mental state. We also demonstrate that a subset of acoustic features has strong discriminative ability in intra-class classifications, such as differentiating between healthy and mild depression levels. To provide a benchmark for comparison, we present detailed experimental results and visual decision processes of depression assessment models. The influence of each acoustic feature is calculated and listed in descending order, providing new insights for physicians to focus on distinguishing depression severity among patients.

The paper is structured as follows. Section II introduces previous corpora, mainly in English. We compare these corpora with our dataset and highlight their advantages and potential shortcomings. We propose our improved data collection strategy in Section III. We present our data preprocessing approach in Section IV. Section V introduces our baseline models and shows the baseline results. Finally, we summarize our findings in Section VI. Researchers interested in this dataset can email the corresponding author to apply for access by signing an end-user license agreement (EULA).

II. RELATED WORK

Most existing automated depression detection approaches utilize supervised learning methods, which were trained using numerous recordings labelled on different depression scales. Therefore, the generalizability of the resulting models relies heavily on the various elements that constitute the dataset. This section will discuss two key elements: data collection methods and depression assessment instruments.

A. Data collection methods

Data collection methods play a crucial role in impacting model performance. Researchers must consider an appropriate context in which the subjects responses are observed. To date, two main types of contexts have been used in collecting depression datasets: social network - an open platform for individuals to share their thoughts, such as scraping social networks to construct depression-related corpora [15], [16]; and spontaneous behaviour - interviewees naturally interact with interviewers or machines, for example, chatting with a chatbot [17], [18]. It is important to note that the choice of data collection method can affect the quality and generalizability of the dataset, and researchers should carefully consider which method is most appropriate for their study.

1) Social network platforms: Datasets for research on affective computing have been well-studied from different perspectives. Herein, we will examine a series of datasets and corresponding data collection strategies. Numerous corpora suitable for diagnosing depression have been collected in low-noise environments and with limited topics. However, these conditions are not representative of the real world, and models trained on such datasets may not perform well when applied to recordings made in uncontrolled settings. In contrast, many researchers have had to perform feature extraction and design application-specific machine learning strategies due to data scarcity. To address the data shortage challenges, Rajput et al. proposed collecting corpora from online forums that focus on mental disease discussions [19], Pirina et al. investigated the influence of the quality of training data [20], De Choudhury et al. and other researchers proposed to collect, analyze, and summarize data from social media platforms, where valuable patterns and information can be detected [21]–[23]. Collecting data from online forums can also significantly reduce difficulties with obtaining sufficient data from healthy individuals. The healthy control group can be sampled from other online communities unrelated to depression. With their vast inflow of user-generated content, social media platforms effectively capture depressive behavioural cues relevant to an individual’s emotional state or mental disorder. However, it is important to note that user-generated content from online forums can be misleading for machine learning models. For example, patients who avoid visiting clinics due to fear of mental disorder-related stigma may also avoid discussing depression online.

2) Interviews under controlled conditions: Many researchers are recruiting volunteers and recording their responses during interviews or free discussions as a method of data collection due to the limitations of social media. The widespread use of smartphones has enabled the emergence of this new strategy to efficiently recruit a diverse sample of participants and collect large amounts of data. Examples of datasets that have adopted this approach include the SEMAINE dataset [24],
which includes audio and video recordings of 150 participants and the Affectiva-MIT Facial Expression Dataset (AM-FED) [25], which consists of labelled spontaneous facial recordings collected over the internet, including 242 video recordings and labels of 10 symmetrical and 4 asymmetrical action units (AU), head movements, smiles, feature tracker confidence, as well as gender and facial landmarks. Dhall et al. proposed a dataset including 4886 images collected in real-world situations with a label on happiness intensity. Moore et al. proposed recording the voice of each subject while reading a short story [26]. Yingthawornsuk et al. also provided a solution to acquiring the recording: two-part interviews between participants and clinicians consisted of an audio recording session. Then the participants were asked to read a selected section of a book [27]. Cohn et al. proposed obtaining facial images from clinical patients in a study of 57 participants [28]. The facial activities and voices of the participants were recorded simultaneously during the interview session with the participants’ permission.

Another issue with the clinical interview data collection strategy is the high cost involved. Gratch et al. proposed an automated interview platform that utilizes an animated virtual interviewer to make patients feel as comfortable as possible [17]. The virtual interviewer can be fully automated or controlled by an operator, which significantly reduces labour costs for data collection. Still, the stringent semi-structured interview process for each patient may be problematic, especially in cases where the patient is unwilling to answer a question. In that case, the virtual interviewer can only proceed to the next question, resulting in patients providing only a few words or nonverbal responses. The resulting datasets may not contain enough information to assess the emotional and mental health state of the patient accurately.

B. Depression assessment instrument

Assessment of depression is challenging due to ongoing research on its pathology [39]. The Diagnostic and Statistical Manual of Mental Disorders (DSM), developed by the American Psychiatric Association, provides the most commonly used set of criterion for diagnosing mental disorders. The DSM aims to provide standard criteria for identifying mental disorders based on observed symptoms such as psychomotor retardation and diminished concentration. The Hamilton Rating Scale for Depression (HAMD) and Beck Depression Index (BDI) are also widely used assessment tools [40], [41]. HAMD is a clinician-administered depression scale and is considered the gold standard assessment tool, while BDI is a self-reported questionnaire. Investigations using both HAMD and BDI have led to the development of new depression scales such as the MADRS [13], Quick Inventory of Depressive Symptomatology (QIDS) [42] and the 9-item Patient Health Questionnaire (PHQ-9) [43]. Previous research has reported that MADRS has higher reliability statistics than QIDS and PHQ-9 [44], [45].

C. The existing corpora

We reviewed previous articles that reported dyadic interview recordings that were annotated based on clinician and outpatient interactions. Approximately half of these datasets were labelled with a self-reported depression rating scale, recorded in controlled conditions, and produced in English. Controlled condition refers to the standardized task and procedures that were used during interviews. Specifically, in previous studies, researchers asked the interviewees to perform tasks such as reading a fixed paragraph, sustained vowels, and memory recalling. This allows investigators to control for variability in responses and simplify the problem. In comparing our dataset to others, we believe that our approach allows us to enable collection of more natural responses from participants. While previous studies have used structured interviews, our methodology allowed for more flexibility in the conversation since participants were able to choose to continue or change the topic as they wished. This approach can yield more spontaneous and authentic responses from the participants, which is important for accurately diagnosing depression. Further, the prevalence of one language (English) in these datasets limits their usability for cross-cultural studies of depression. Table I compares existing data from social networks and clinical interviews.

III. Data Collection

The main goal of this study was to collect high-quality responses from subjects participating in clinical depression interviews. Previous research has found that spontaneous speech is more effective than reading speech in depression classification [46], [47]. In addition, this study aims to examine subjects’ emotional responses to physicians’ questions. Therefore, the data collection protocol, related experiments, and data preprocessing procedures have been designed to detect and evaluate depression and depression levels.

A. Participants

Participants (n=113) were recruited for a psychology study with informed consent and ages ranging from 15 to 65. Participants were required to be native Mandarin speakers with at least primary education. To ensure that our findings were applicable to a broader population, we carefully selected a representative sample of individuals with depression. We also took recommendations from clinicians into account which excluded individuals with a history of antidepressant medication or mental disorders. Participants who were diagnosed with depression had no other mental or medical conditions were eligible. Verbal consent and signed forms were obtained, allowing data processing and distribution with removed patient identification. The study was conducted in Wenzhou, China, with in-person interviews taking place in a confidential private room that was pre-arranged for the purpose of protecting patients’ privacy. Although the interviews were conducted in a private room, we did not use noise-cancelling equipment or impose any restrictions on the topics discussed during the interviews. Our goal was to capture a range of natural variations that might occur in real-world settings, thus ensuring the authenticity and generalizability of our dataset, which is important because noise levels in public spaces are not controlled. If the model is trained on noise-cancelling data,
its performance may not be as robust in real-world settings. We ensured the standardization of our data collection process in several ways. First, we employed experienced physicians to conduct all the interviews. A total of four attending physicians, each with a minimum of five years of experience, are involved. We also utilized the MADRS questionnaire, a reliable and well-established tool with high inter-rater consistency and reliability. We also followed well-established standardized protocols for conducting interviews, recording data, and managing data quality. Before starting the full-scale data collection, we conducted pilot testing, including random selection of some outpatients to conduct depression interviews on to ensure the data collection process is feasible, reliable and standardized. In addition, we developed a data management plan that outlined procedures for storing, protecting and sharing data to ensure that data quality and privacy were maintained throughout the data collection process. Clinicians were not aware of the mental health condition of the examined subject in advance. Interviews were conducted in Mandarin and lasted 5-10 minutes [13]. Clinicians had the flexibility to adjust the order of questions within the MADRS questionnaire, and they also allowed patients to discuss other related topics. Our approach allowed us to gather more comprehensive and individualized data on each participant. Audios were recorded in real-time at a 48 kHz sampling rate, 128 kbps bitrate, and mono-channel MP3 format. The study was approved by the ethics committee of Wenzhou Kangning Hospital (No. AF/SQ-02/01.0).

### B. Procedure

After obtaining verbal consent and a signed form for recording, the MADRS questionnaire interview was conducted by clinicians in Chinese. The MADRS, consisting of 10 items rated on a 6-point scale, evaluates core depression symptoms, with a maximum possible score of 60 points. Scores between 7-19 indicate mild depression, 20-34 indicate moderate depression, and scores above 34 indicate severe depression [13]. The questionnaire focused on the ten critical symptoms in Table III. During the experiment, participants were asked questions by a clinician about their mental health. The order of the questions may have varied at the clinician’s discretion. In some cases, additional questions were asked for more information based on the clinician’s judgement and experience, as long as the question was still relevant to the previous topic, and the participant was willing to discuss it. The clinician was also allowed to adjust the initial questions to put the participant at ease. At the end of the interview, the clinician helped the participant relax from any distress they may have experienced. Experienced clinicians conducted the interviews to minimize any further impact on the participants’ mental health. Our goal was to elicit verbal and non-verbal cues of depression from the participants.

### C. Dataset statistics

In our study, we interviewed 113 participants, 52 of whom were healthy, and 61 were depressed patients. The interview audios were an average of 364.40 seconds in length (st. dev = 257.66 seconds). For the control group, the audio files were an average of 164.53 seconds in length (st. dev = 101.88 seconds), and the average sentence word count is 6.14 (st. dev = 6.44). For the depressed patient group, the audio files were have an average of 535.70 seconds in length (st. dev = 224.78 seconds), and the average sentence word count is 6.41 (st. dev = 5.89). Patient demographics are illustrated in Table II. The supplementary material shows the interview sample in Table S1 and S2. Before further analysis, a balanced dataset was built by random sampling. For binary depression detection, positive and negative samples should be approximately equal. For multiclass depression level prediction, the distribution of severity levels should be balanced. In our dataset, 52 and 61 participants were in the healthy and depressed patient groups, respectively. Fig. II(a) shows the distribution of the depression levels in our dataset. Fig. II(b) and Fig. II(d) illustrate the distribution of the audio duration in healthy and depressive groups. The average audio duration for the depressive population was shorter than that of the healthy population.

### Table I: A Comparative Study of the Proposed Dataset and Datasets Employed in the Reviewed Studies for Depression Detection

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Population/Healthy vs depressive</th>
<th>Collection protocol</th>
<th>Language Label</th>
<th>Criteria</th>
<th>Research purposes</th>
<th>Video resolution</th>
<th>Modality</th>
<th>Controlled condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xu et al. 2020 [29]</td>
<td>750 Social media</td>
<td>English Self-report</td>
<td>-</td>
<td>Detection</td>
<td>-</td>
<td>Text</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>BlackDog [31]</td>
<td>130 (70/60)</td>
<td>English Clinical assessment</td>
<td>DSM-IV: HAMD&gt;15</td>
<td>Detection</td>
<td>-</td>
<td>Visual &amp; Audio</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>OBI [34]</td>
<td>31 (4/4)</td>
<td>Interpersonal</td>
<td>English</td>
<td>-</td>
<td>Detection</td>
<td>-</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>ORYGEN [35]</td>
<td>30 (15/15)</td>
<td>Interpersonal</td>
<td>English Clinical assessment</td>
<td>-</td>
<td>Detection</td>
<td>-</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>CHE-MIB [36]</td>
<td>26 (13/13)</td>
<td>Interpersonal &amp; Virtual agent</td>
<td>Chinese Clinical assessment</td>
<td>HAMD=15</td>
<td>Detection</td>
<td>640x480</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td>EMORY [37]</td>
<td>7 (4/7)</td>
<td>Interpersonal</td>
<td>English Clinical assessment</td>
<td>HAMD=15</td>
<td>Recovery</td>
<td>-</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>CMDC [38]</td>
<td>78 (26/52)</td>
<td>Interpersonal</td>
<td>Chinese Clinical assessment</td>
<td>HAMD=17 or PHQ-9&gt;9</td>
<td>Detection &amp; Severity</td>
<td>-</td>
<td>Visual &amp; Audio &amp; Text</td>
<td>No</td>
</tr>
</tbody>
</table>

### Table II: The Questionnaire Used During Interview

<table>
<thead>
<tr>
<th>Symptom</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apparent sadness</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>Reported sadness</td>
<td>How is everything going?</td>
</tr>
<tr>
<td>Inner tension</td>
<td>Have you ever been feeling nervous and scared for no reason?</td>
</tr>
<tr>
<td>Reduced sleep</td>
<td>How do you sleep recently?</td>
</tr>
<tr>
<td>Reduced appetite</td>
<td>How do you eat recently?</td>
</tr>
<tr>
<td>Concentration difficulties</td>
<td>Can you stay focused?</td>
</tr>
<tr>
<td>Lassitude</td>
<td>Do you feel like you don’t want to do anything?</td>
</tr>
<tr>
<td>Inability to feel</td>
<td>Do you feel that everything has nothing to do with you?</td>
</tr>
<tr>
<td>Pessimistic thoughts</td>
<td>Do you feel inferior or self-blaming?</td>
</tr>
<tr>
<td>Suicidal thoughts</td>
<td>Have you ever thought of self-harm or suicide?</td>
</tr>
</tbody>
</table>
significantly longer than that of the healthy. The distribution of the utterance length is shown in Fig. 4 (c) and Fig. 4 (e). Compared to the distribution of the audio duration, the number of words in a sentence for the control and experiment groups was not significantly different (p > 0.05). To identify patterns between the healthy and depressive groups, we generated word clouds and showed frequently used words in a larger font in Fig. 4 (f) and Fig. 4 (g). The word cloud of the depressive group reveals that these individuals are more likely to use negative words such as ‘difficult to fall asleep,’ ‘being in a bad mood,’ and ‘not good’ during the interview.

IV. DATA PROCESSING

In this section, we describe the preprocessing procedure for interview recordings, using prosodic and acoustic features, frequency-based features and pre-trained word embeddings.

A. Audio recordings preprocessing

1) Audio transcription: The iFlyTek API was used to transcribe audio recordings, which were then reviewed by research assistants majoring in psychiatry. Before starting the automatic transcription, raw audio files larger than 10 MB were divided into smaller data blocks as required by the transcription algorithm. The transcript blocks were merged sequentially using each block’s unique ID to create the final transcript. The raw transcription results were in JavaScript Object Notation (JSON) format, containing various fields such as the timestamp of a sentence, sentence content, speaker identification, and sentence tokenization. The speaker identification helped to isolate the patients’ responses in the raw audio. The timestamp of a sentence, indicating each sentence’s start and end points, was used to extract the patient’s audio clips and vocal features in later experiments. The tokenization result was used as input for the frequency vectorizer. These JSON objects were parsed using the Python internal JSON package and converted into comma-separated values (CSV) files.

2) Audio feature extraction: We used the COVAREP toolkit to capture the frame-level acoustic features [14]. COVAREP is an open-source feature extraction toolkit commonly used in depression classification studies. We segmented the raw interview audio recordings using the COVAREP toolkit, which allowed us to extract audio features at a rate of 100Hz. To achieve this, we divided the raw recording into 10 millisecond blocks, which is common practice in speech processing [48]–[51]. We then read in each interview recording from the input directory and extract various features for each block in the recording. Specifically, we extract features such as $F_0$, voiced/voiceless (VUV) decision, NAQ, quasi-open quotient (QQQ), H1-H2, peak-slope (PSP), modulation depth quotient (MDQ), relative amplitude quotient (Rd), creaky voice detection, Mel-Cepstral coefficients (MCEPs), and Harmonic Model + Phase Distortion (HMPD) features. By using a 10-millisecond block size and a sampling rate of 100Hz, we believe that we were able to capture the relevant acoustic information at a reasonable computational cost. Detailed descriptions of each audio feature can be found in Table 4.

a) Fundamental frequency $F_0$: The fundamental frequency $F_0$ was investigated from various aspects. Regarding a stationary and periodic signal, the fundamental frequency is given by the inverse of its period. However, since speech signals are non-stationary and time-variant, the position of the vocal tract can change abruptly. Therefore, the starting point of the measurement cannot be ignored as it influences the final measurement. Previous articles have proposed different algorithms to estimate $F_0$ with the attributes of speech signals in the time and spectral domains, while other researchers have proposed exploitation of both spaces. $F_0$ has been described in many previous studies as a biomarker of depression [52]–[54], which exhibited strong discriminating power in distinguishing depression and other mental disorders.

b) Glottal flow features: Compared with fundamental frequency $F_0$, the glottal flow features have received less attention in previous depression and mental disorders studies. Moore et al. demonstrated that several glottal flow features exhibited significant separation between control and healthy groups [26]. Speech production lasts several glottal cycles, with each cycle involving an open phase (O) and a closed phase (C). NAQ [55] and QOQ [56] are two features calculated from the glottal flow, which is given by [57]:

$$\text{NAQ} = \frac{f_{ac}}{d_{\text{peak}} \cdot T_0}$$  (1)

where $d_{\text{peak}}$ is the negative amplitude of the main excitation in the differentiated glottal flow pulse, $f_{ac}$ is the peak amplitude of the glottal flow pulse, $T_0$ is the length of the glottal pulse period. QOQ is calculated by amplitude measurements of the glottal flow pulse. The quasi-open period is measured by finding the peak in the glottal flow and the time points prior to the peak that descends below 50% of the amplitude. The duration between the two time points is divided by the local glottal period to determine QOQ [57]. Besides QOQ and NAQ, other glottal features have been commonly used in previous articles on automatic depression detection [57], [58].

B. Transcripts preprocessing

Our proposed dataset includes 113 transcripts in comma-separated value files, with five fields per transcript: “bg,” “ed,” “speaker,” “value,” and “words list.” The “bg” and “ed” fields indicate the start and end of one sentence captured by the transcription algorithm. The “value” field is the sentence recognized and transcribed by the algorithm, and “words list” field is the sentence tokenization. The “value” and “speaker” fields may contain errors due to environmental noise or a lack of pause between the psychiatrist and patient. After the transcriptions were verified against the audio recordings by research assistants, the sentences in the transcripts were tokenized using Jieba, a Chinese tokenization library. The transcripts were then divided into healthy and depressive groups based on the physicians’ diagnosis after removing stop words such as “if” and “too.”

V. BASELINE RESULTS

We established baseline models and performance metrics to detect depression and classify its severity on our proposed
dataset. We conducted two series of experiments, one using
frequency-based text features and the other using a set of audio
features. Due to the limited size of our dataset, we employed
nested cross-validation over the training set to ensure objective
performance evaluation. In related works (see Table I), most
clinical datasets comprise 100 to 200 data points, due to the
high cost of data collection.

A. Experimental Setting

This study focused on detecting depression and predicting
disease severity. To achieve this, the dataset was split into
independent training and test sets. The training set consisted
of 41 healthy individuals and 49 individuals with depression,
while the test set included 10 healthy individuals and 13

<table>
<thead>
<tr>
<th>Gender</th>
<th>Subjects categorized as depressed</th>
<th>Subject categorized as healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>46</td>
<td>33</td>
</tr>
<tr>
<td>Male</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>Age</td>
<td>Mean: 27.29 Std: 9.45</td>
<td>Mean: 32.70 Std: 6.45</td>
</tr>
<tr>
<td>&lt;=20</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>21-25</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>26-30</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>31-35</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>&gt;35</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td>Marries</td>
<td>36</td>
<td>45</td>
</tr>
<tr>
<td>Divorced</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Academic qualification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary school</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Secondary school</td>
<td>31</td>
<td>9</td>
</tr>
<tr>
<td>Diploma/ Degree</td>
<td>28</td>
<td>23</td>
</tr>
<tr>
<td>Master</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Working status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full time</td>
<td>29</td>
<td>38</td>
</tr>
<tr>
<td>Part time</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Unemployed/ Student</td>
<td>31</td>
<td>12</td>
</tr>
</tbody>
</table>

**TABLE III: Summary of dataset characteristic**

**Fig. 1:** The proposed dataset contains 113 individuals, (a) 51 of whom are healthy and 62 of whom are patients with depression. Of the patients with depression, 9 have mild depression, 34 have moderate depression, and 19 have severe depression. (b) The distribution of audio duration between the healthy and depressive subjects. (c) The distribution of utterance length between the healthy and depressive groups. (d) The distribution of audio duration across four depression levels. (e) The distribution of utterance length across four depression levels. (f), (g) The word cloud of the healthy (above) and the depressive (below) groups. (h), (i) The word cloud in English. Negative words and phrases, such as “difficult to fall asleep”, “bad mood”, etc., are in the word cloud below.
individuals with depression.

To assess the performance of the models in a more challenging scenario, the dataset was further divided into training and test sets based on depression severity. The training set included 41 healthy individuals, 7 with mild depression, 27 with moderate depression, and 15 with severe depression. The test set included 10 healthy individuals, 2 with mild depression, 7 with moderate depression, and 4 with severe depression.

To maintain balance among minority classes, we over-sampled the minority classes in the training set. For each experiment, we fine-tuned the models by conducting a grid search with nested cross-validation to determine the optimal hyper-parameters and then reported the results on the test set. Apart from the baseline experiments, we also assessed the effectiveness of pretrained deep learning models \cite{59} on the proposed dataset. For the deep learning-based experiments, we labelled each frame of audio and text features with the subjects’ diagnostic results. The dataset was then split into training, validation, and test sets with an 8:1:1 ratio, similar to the method described in \cite{59}.

### B. Baseline text models

To identify depression, we categorized transcripts into two groups - healthy and depressive - as assessed by psychiatrists. We also classified transcripts into four groups based on their MADRS scores to determine level of depression. All transcripts were encoded in “GBK” for easy analysis, and commas separated all fields. In the text-based experiments, we used uni-gram \cite{60}, bi-gram \cite{61}, Term Frequency-Inverse Document Frequency (TF-IDF) \cite{62}, and pre-trained word embeddings to represent transcripts.

Uni-gram refers to a single word or token that appears in a document. It is the simplest and most commonly used form of text representation in natural language processing (NLP). Uni-gram count is the frequency of a particular word or token within a single document. Bi-gram refers to two consecutive words or tokens that appear in a document. Bi-gram count is the frequency of two consecutive words that appear together in a document. TF-IDF is a statistical measure that evaluates the importance of a word or token in a document. It takes into account the frequency of a word within a single document (term frequency) and across all documents in a corpus (inverse document frequency). It is used to determine the relative importance of words or tokens in a document and is widely used for information retrieval and text mining.

The count vectorizer builds a vocabulary by scanning all transcripts and transforming each document into a matrix of token counts. The count vectorizer builds a vocabulary by scanning all transcripts and transforming each document into a matrix of token counts. Let $x = (x_1, x_2, \ldots, x_n), x \in R^{n \times n}$, be a vector of token frequencies, where each column vector $x_i$ represents token frequencies of subject $S_i$, where $s$ is the number of subjects, $n$ is the size of the vocabulary.

After generating frequency-based text features, we trained multinomial Bayes classifiers to identify depression and predict its severity for each subject. We selected the multinomial classifier because of its ability to classify numerical features and used it to calculate the probability of depression presence level based on each subject’s feature vector, i.e.

$$P(C = c_i \mid X = x_i) = \frac{P(X = x_i \mid C = c_i) P(C = c_i)}{P(X = x_i)}$$

(2)

The probability of any given example can be assigned a class $C_i$, which is given by:

$$P(X = x_i \mid C = c_i) P(C = c_i)$$

(3)

Therefore, the Bayes classifier finds the maximum a posterior probability (MAP) given any example $x$, i.e.

$$h^*(x) = \arg \max_i P(X = x_i \mid C = c_i) P(C = c_i)$$

(4)

However, $x_i$ is a high-dimensional feature vector, resulting in difficulties directly computing the term $P(X = x_i \mid C = c_i)$. An approximation is adopted to reduce the computation cost, such as using the assumption that features are conditionally independent given the class $C_i$, i.e.

$$f_i(x) = \prod_{j=1}^{m} P(X_j = x_j \mid C = c_i) P(C = c_i)$$

(5)

The set of conditional probabilities in Equation 5 can prove unreliable when a word is missing from the training set; regardless of its label, a zero product of conditional probability is the result. Avoiding zero-conditional probabilities is accomplished by adopting a smoothed conditional probability instead of directly computing the conditional probability $p(x \mid y, c_i)$, which is given by:

$$P(x \mid y, c_i) = \frac{P(x, y, c_i) + \alpha \cdot P(x)}{P(y, c_i) + \alpha}$$

(6)

1) Binary classification text model (depression vs. healthy):

We trained and evaluated multinomial Bayes models using nested cross-validation on the training set collected for this study. During the hyperparameter fine-tuning phase, we optimized the classifier parameters to achieve the highest $F_1$ score. To eliminate unreliable estimates and zero-conditional probabilities, we optimized the parameter $\alpha$ for the multinomial Bayes classifiers. We expected that the conditional probability of a given word that only exists in the test set is close to zero; therefore, $\alpha$ varied in the range $(10^k \, \text{with} \, k = 0, \ldots, 3)$. The best parameter $\alpha$ was determined to be 1.0 after cross-validation, and the best micro average $F_1$ score was 0.85 in

\begin{table}[h]
\centering
\caption{COVAREP Spectral and Cepstral Feature Set}
\begin{tabular}{|l|c|}
\hline
\textbf{Voicing based} & \textbf{Group} \\
\hline
M1 & Spectral \\
VUV & Spectral \\
PSP & Spectral \\
\hline
\textbf{Glottal source-based} & \textbf{Group} \\
Normalised amplitude quotient (NAQ) & Spectral \\
H1, H2 & Spectral \\
Parabolic Spectral Parameter (PSP) & Spectral \\
\hline
\textbf{Spectral envelope-based} & \textbf{Group} \\
\text{Mel} cepstral coefficients (MCEP1-MCEP6) & Cepstral \\
Harmonic model and phase distortion mean (HMPD1,HMPD2) & Cepstral \\
Harmonic model and phase distortion deviation (HMPDD1,HMPDD2) & Cepstral \\
\hline
\textbf{Wavelet-based} & \textbf{Group} \\
Maxima Dispersion Quotient (MDQ) & Spectral \\
Peak slope & Spectral \\
\hline
\end{tabular}
\end{table}
the cross-validation and 0.91 in the test set. The details of other metrics are listed in Table [V].

2) Depression level classification: To investigate whether the severity of depression was related to our extracted text features, we used a multinomial Bayes model for depression severity classification. Since our dataset was limited in size, we conducted nested cross-validation to train the model and evaluate its performance. As mentioned in the second to last paragraph in Section [V-A], the training set included 41 healthy individuals, 7 with mild depression, 27 with moderate depression, and 15 with severe depression, while the test set included 10 healthy individuals, 2 with mild depression, 7 with moderate depression, and 4 with severe depression. By merging the “none” and “mild” classes and the “moderate” and “severe” classes, we were able to create a more balanced training set with 48 individuals in the none&mild class and 42 in the moderate&severe class. This balanced dataset allowed us to train our machine learning model more effectively and produce more accurate results. We chose to use the F1 score as our evaluation metric because it is more sensitive to data distribution. In healthcare datasets, there are often more patients than healthy individuals, making it important to choose an evaluation metric that is appropriate for imbalanced datasets. The F1 score takes into account both precision and recall, making it a suitable metric for evaluating the performance of our machine learning model on imbalanced data. To optimize the multinomial Bayes classifier, we varied the parameter $\alpha$ in the range of $10^k$ where $k$ was set to 0, 1, 2, or 3. The best value for $\alpha$ was determined to be 100 through nested cross-validation. The best micro average $F_1$ score was 0.58 in the cross-validation, and the same score was obtained on the test set. Details of other metrics can be found in Table [VI].

C. Baseline audio models

We used COVAREP to extract audio features from each participant’s interview recording at a rate of 10 milliseconds. The interviews lasted between 5 to 10 minutes, resulting in a variable number of frames extracted from the recordings. This caused difficulties in batch processing. To overcome this issue, we employed a histogram-based processing technique known as the “Neighborhood top-N elements method” to transform the variable-length audio feature frames into fixed-length ones.

1) Neighborhood top-N elements method: To obtain the audio feature frames with a fixed-length, we computed the histogram of each audio feature to determine its global distribution during the interview. The top-N most frequent elements in the histogram represent the audio feature, and we used the left-endpoints of these elements. To compute the histogram, we needed to determine the number of bins for each audio feature. If the number of bins was too small, most of the entries were grouped into the same bin. Conversely, if the number of bins was too high, only a few entries would be in each bin. We avoided these situations as they may not accurately describe the statistical characteristics of the audio features. To determine the number of bins, we used the Freedman-Diaconis rule, which calculates the bin width to minimize the difference between the area under the empirical data distribution and the theoretical data distribution. [63]. The hyperparameter $N$ for each audio feature was determined in advance using nested cross-validation. Specifically, we tested values of $N$ in the range [5, 10, 15, 20, 25, 30, 35] and recorded the value that resulted in the best cross-validation score on the training partition of the original dataset.

2) Binary classification audio model (depression vs. healthy): XGBoost is an open-source research project that implements a tree-based gradient-boosting algorithm. The XGBoost model has many useful features including: it is an ensemble learning method, which decreases the bias of the model; and it is a tree-based model with high interpretability, which aids in determining the feature’s importance in making an inference. These models also offer a good trade-off between computation cost and accuracy. Tree-based boosting algorithm methods solve many machine learning problems efficiently and accurately, making such methods good candidates for providing baseline results within our dataset. To train our XGBoost classifiers, we created a separate model for each audio feature. We excluded certain features, such as HMPDM_0 to HMPDM_3, since they remained constant throughout the interview. To facilitate our final decision, we applied a majority voting algorithm to the output of each classifier. To optimize our models, we fine-tuned the parameters by maximizing the $F_1$ score, which we deemed equally important for precision and recall. For each XGBoost classifier, we tuned several parameters, including the learning rate, max depth of the tree, and number of estimators. To identify the optimal hyperparameters, we conducted a grid search on the training set, selecting models with high precision and recall. In the nested cross-validation, we achieved a best micro average $F_1$ score of 0.81, which improved to 0.87 on the test set. Further details on other metrics can be found in Table VII.

3) Depression level classification: In our investigation of the relationship between depression severity and audio features, we trained depression severity prediction models using the top-N elements method. This method transformed variable-length audio features into fixed-length vectors, which were then used in our analysis. We trained and evaluated each model on both the training and validation set, testing different parameters to optimize performance. The highest performing model, which we chose for our study, achieved an F1 score of 0.52 in cross-validation and 0.55 on the test set. Results of the fine-tuned baseline models can be found in Table [VIII].

4) Multimodality baseline models: To enhance the model’s ability to assess depression, we employed late fusion to combine the outputs of the acoustic and semantic models. Our multimodality baseline models produced an output through a linear combination of the acoustic and semantic model outputs. In Table [V] and Table VII, the depression detection accuracy of the acoustic-only and semantic-only models were 0.82 and 0.81, respectively. For depression-level classification, the accuracy of the semantic-only and acoustic-only models were 0.62 and 0.61, respectively, as shown in Table VI and Table VIII. During cross-validation and on the test set, our multimodality depression detection model (accuracy=0.86, see Table IX) and multimodality depression-level classification model (accuracy=0.63, see Table X) produced fewer errors...
than the acoustic-only and semantic-only models.

VI. FEATURE STATISTICS

A. Audio features

Our dataset analyzed recordings from 113 clinically supervised participants, resulting in two different comparisons: inter-condition and intra-condition comparisons. Inter-condition comparisons evaluate differences in audio features between healthy and depressive groups, such as whether participants from the control (healthy) and experimental (depressive) groups differ in vocal fundamental frequency ($F_0$) as processed by the top-$N$ elements method. Intra-condition comparisons evaluate the variability of a patients' audio features relative to their severity of depression. This second comparison

<table>
<thead>
<tr>
<th>Feature</th>
<th>State</th>
<th>Precision CV</th>
<th>Test</th>
<th>Recall CV</th>
<th>Test</th>
<th>F1 score CV</th>
<th>Test</th>
<th>Accuracy CV</th>
<th>Test</th>
<th>Cohen Kappa CV</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uni-gram</td>
<td>Healthy</td>
<td>0.94 ± 0.07</td>
<td>1.00</td>
<td>0.73 ± 0.09</td>
<td>0.80</td>
<td>0.82 ± 0.06</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Depressive</td>
<td>0.81 ± 0.05</td>
<td>0.87</td>
<td>0.96 ± 0.07</td>
<td>1.00</td>
<td>0.88 ± 0.05</td>
<td>0.93</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>0.87 ± 0.06</td>
<td>0.92</td>
<td>0.86 ± 0.05</td>
<td>0.91</td>
<td>0.85 ± 0.03</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bi-gram</td>
<td>Healthy</td>
<td>0.96 ± 0.09</td>
<td>1.00</td>
<td>0.56 ± 0.02</td>
<td>0.70</td>
<td>0.71 ± 0.03</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Depressive</td>
<td>0.73 ± 0.04</td>
<td>0.81</td>
<td>0.98 ± 0.04</td>
<td>1.00</td>
<td>0.83 ± 0.06</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>0.83 ± 0.09</td>
<td>0.89</td>
<td>0.79 ± 0.05</td>
<td>0.87</td>
<td>0.78 ± 0.09</td>
<td>0.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TF-IDF</td>
<td>Healthy</td>
<td>1.00</td>
<td>1.00</td>
<td>0.59 ± 0.12</td>
<td>0.70</td>
<td>0.74 ± 0.11</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Depressive</td>
<td>0.74 ± 0.08</td>
<td>0.81</td>
<td>1.00</td>
<td>0.85 ± 0.06</td>
<td>0.90</td>
<td>0.81 ± 0.05</td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>0.86 ± 0.07</td>
<td>0.89</td>
<td>0.81 ± 0.03</td>
<td>0.87</td>
<td>0.80 ± 0.06</td>
<td>0.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Word embeddings [59]</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>0.96</td>
<td>-</td>
<td>0.98</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>0.95</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>-</td>
<td>0.98</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>State</th>
<th>Precision CV</th>
<th>Test</th>
<th>Recall CV</th>
<th>Test</th>
<th>F1 score CV</th>
<th>Test</th>
<th>Accuracy CV</th>
<th>Test</th>
<th>Cohen Kappa CV</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uni-gram</td>
<td>None &amp; Mild</td>
<td>0.91 ± 0.04</td>
<td>0.73</td>
<td>0.76 ± 0.08</td>
<td>0.80</td>
<td>0.83 ± 0.07</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate &amp; Severe</td>
<td>0.43 ± 0.06</td>
<td>0.50</td>
<td>0.74 ± 0.08</td>
<td>0.86</td>
<td>0.54 ± 0.08</td>
<td>0.66</td>
<td>0.58 ± 0.06</td>
<td>0.61</td>
<td>0.36 ± 0.02</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>0.57 ± 0.07</td>
<td>0.47</td>
<td>0.58 ± 0.06</td>
<td>0.61</td>
<td>0.55 ± 0.09</td>
<td>0.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bi-gram</td>
<td>None &amp; Mild</td>
<td>0.89 ± 0.08</td>
<td>1</td>
<td>0.80 ± 0.05</td>
<td>0.80</td>
<td>0.85 ± 0.04</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate &amp; Severe</td>
<td>0.46 ± 0.05</td>
<td>0.47</td>
<td>0.81 ± 0.07</td>
<td>1</td>
<td>0.59 ± 0.03</td>
<td>0.64</td>
<td>0.62 ± 0.04</td>
<td>0.65</td>
<td>0.41 ± 0.02</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>0.59 ± 0.03</td>
<td>0.58</td>
<td>0.62 ± 0.05</td>
<td>0.65</td>
<td>0.58 ± 0.04</td>
<td>0.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TF-IDF</td>
<td>None &amp; Mild</td>
<td>0.79 ± 0.05</td>
<td>0.80</td>
<td>0.90 ± 0.03</td>
<td>0.80</td>
<td>0.84 ± 0.04</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate &amp; Severe</td>
<td>0.51 ± 0.04</td>
<td>0.38</td>
<td>0.81 ± 0.05</td>
<td>0.71</td>
<td>0.63 ± 0.05</td>
<td>0.50</td>
<td>0.66 ± 0.05</td>
<td>0.57</td>
<td>0.44 ± 0.03</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>0.51 ± 0.02</td>
<td>0.46</td>
<td>0.66 ± 0.04</td>
<td>0.57</td>
<td>0.57 ± 0.02</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Word embeddings [59]</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>0.96</td>
<td>-</td>
<td>0.98</td>
<td>-</td>
<td>0.85</td>
<td>-</td>
<td>0.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>State</th>
<th>Precision CV</th>
<th>Test</th>
<th>Recall CV</th>
<th>Test</th>
<th>F1 score CV</th>
<th>Test</th>
<th>Accuracy CV</th>
<th>Test</th>
<th>Cohen Kappa CV</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>Healthy</td>
<td>0.83 ± 0.04</td>
<td>0.89</td>
<td>0.73 ± 0.05</td>
<td>0.80</td>
<td>0.78 ± 0.02</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Depressive</td>
<td>0.80 ± 0.04</td>
<td>0.86</td>
<td>0.88 ± 0.04</td>
<td>0.92</td>
<td>0.83 ± 0.05</td>
<td>0.89</td>
<td>0.81 ± 0.04</td>
<td>0.87</td>
<td>0.62 ± 0.06</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>0.81 ± 0.07</td>
<td>0.87</td>
<td>0.81 ± 0.05</td>
<td>0.87</td>
<td>0.81 ± 0.04</td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>Healthy</td>
<td>0.75 ± 0.04</td>
<td>0.78</td>
<td>0.66 ± 0.03</td>
<td>0.70</td>
<td>0.70 ± 0.05</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Depressive</td>
<td>0.74 ± 0.03</td>
<td>0.79</td>
<td>0.82 ± 0.04</td>
<td>0.85</td>
<td>0.78 ± 0.05</td>
<td>0.81</td>
<td>0.74 ± 0.05</td>
<td>0.78</td>
<td>0.48 ± 0.02</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>0.74 ± 0.03</td>
<td>0.78</td>
<td>0.74 ± 0.02</td>
<td>0.78</td>
<td>0.74 ± 0.04</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BLSTM+T-CNN [59]</td>
<td>Healthy</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>0.96</td>
<td>-</td>
<td>0.98</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Depressive</td>
<td>-</td>
<td>0.95</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>-</td>
<td>0.98</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
<td>0.97</td>
</tr>
</tbody>
</table>
is crucial because it offers a new way of understanding whether an individual's depression severity changes by focusing on specific audio features. In this section, we focused on comparing the distribution of three specific features, namely $F_0$, MCEP0, and HMPDM_17. We conducted statistical tests to determine if there were significant differences in the distributions of these features between the depressed and non-depressed groups, as well as between different levels of depression. Regarding the other 71 features, we found that their distributions were not significantly different between the two groups, and therefore did not include them in our comparison. However, we want to emphasize that these features may still be useful for future research and could potentially provide further insights into the relationship between speech and depression.

1) Vocal fundamental frequency ($F_0$): $F_0$ is one of the significant acoustic variables correlating to the pitch; $F_0$ is determined by the vibration frequency of the vocal fold and is used to describe the periodicity of the speech. Our analysis found that the inter-condition effect is present for female participants for $F_0$. The median $F_0$ of the healthy control group is lower compared to participants from the depressive group, as shown in Fig. 2. This is in line with the conclusion reached by Mundt et al. that the healthy control group had a lower $F_0$ than the depressive group [64]. The variances of $F_0$ between the two groups were compared using a Welch t-test, and the variance of $F_0$ of the healthy group was found to be significantly greater than that of the depressive (p<0.01). However, $F_0$ was not a significant audio feature in male participants.

2) Mel-Cepstrum Coefficient (MCEP): MCEP was included in our model as it has been effectively demonstrated in the characterization of speech content [65]–[67]. In our research, we conducted a Welch t-test to determine if MCEP audio features differ significantly with depression presence and severity.
For the binary classification task (depressive vs. healthy), we identified MCEP values that were significantly different between the healthy and depressive groups. The box plots in Fig. 3 and 4 confirm that MCEP_0 can be used in both male and female groups as a criterion to distinguish potentially depressed patients. However, some high-order MCEPs (such as MCEP_8, MCEP_13 and MCEP_18) had overlapping values between the healthy and depressive subjects. MCEP_0 was significantly different in the healthy, mild, moderate and severe depression groups, as shown in Fig. 3 and 4. Therefore, MCEP_0 may be a gender-independent factor in distinguishing depression presence and severity.

3) Harmonic model and phase distortion mean (HMPDM):
Several reports have shown that HMPDMs can be used to predict depression presence and severities [68]–[71]. In our research, we conducted a Welch t-test to determine if HMPDM values in the healthy and depressive groups were significantly different. The significance levels were set at 1% for HMPDM audio features. For the binary classification (depressive vs. healthy), the higher-order HMPDMs, such as the HMPDM_17, were found to be significantly different between the healthy and depressive subjects (see Fig. 5), suggesting that HMPDMs may play a key role in predicting depression. Additionally, the variance of HMPDM_17 increased in participants suffering from depression, while the median of the HMPDM_17 was higher in healthy subjects. In depression severity classification, HMPDM_17 of female participants was a reliable indicator for predicting depression levels. For example, Fig. 6 shows that the healthy group has a higher median of HMPDM_17. Further
investigation is needed to fully understand the role of each audio feature in depression detection and level classification.

### B. Impact of audio features during inference

Recent studies have made significant progress in using machine learning models in combination with mental healthcare to assist in depression diagnosis [72]–[75]. However, providing doctors with a clear and natural explanation of the criteria used in a prediction can be challenging. For example, a numeric probability of depression is helpful, but it may not provide enough information for a doctor to understand how the prediction was made. To provide a more clinically meaningful explanation, using audio features such as $F_0$, MCEP, and HMPDM may be more informative. Generally, explaining how a prediction was made limits the model we can use, but we chose to adopt the Shapley Additive exPlanations (SHAP) proposed by Lundberg et al. [76]. This approach allows us to understand the contribution of each audio feature to the prediction by comparing the output of the model when a feature is included or excluded. However, it is important to note that the feature contribution does not demonstrate causality and does not represent a final diagnosis of depression. It enables doctors to make more informed diagnoses by understanding which audio features contribute more to the generated depression prediction.

To demonstrate the reliability of the predictions made by the model and gain further insight into factors that affect depression diagnosis, we present the graphical contribution of audio features to the prediction process. Our model outputs depression probability and its explanations, which shows a series of features that increased (red) and decreased (blue) the depression risk. Based on professional diagnosis by clinicians, we divided the dataset into two categories: healthy and depressive. The audio features were extracted and processed using the method described in Section V-C. The original dataset was split into training (80%) and test (20%) sets. We trained an XGBoost binary classification model with the optimal parameters obtained in Section V-C. The output of the binary classifier provided the depression probability of the participant. An explanation of our model represents the contributions of interpretable groups of audio features. These contributions explain how the model makes a prediction, making it possible for psychiatrists to reach a final diagnosis. In section VI-A, we only investigated the difference between each preprocessed audio feature (processed by the neighbourhoud top-$N$ elements method) in the healthy and depressive groups. Without a meaningful explanation, the output probability of the model may be difficult to interpret. By presenting the depression probability as a cumulative process, the reason for the prediction becomes clearer.

The increase in the depression probability of test examples shown in Fig. 7-10 is driven by audio features. The probability explanation bar in Fig. 7-10 has red features that push the probability higher (to the right) and blue features that push the probability lower (to the left). The magnitude of their contribution sorts audio features, and the features with the higher contributions are labelled. Through this representation, we can conclude that most audio features have a small impact, while a few are responsible for driving the probability of the depression features. Instead of feeding the model with important features, we allow the model to select the features it believes to be effective, meaning that the model may select unimportant features, which makes it possible for psychiatrists to interpret the features better.

### VII. Conclusion

Open datasets are valuable for both the research and clinical communities. Collecting and annotating clinical interviews with professional diagnoses is labour-intensive and requires expertise in psychology. Authentic depression interview
datasets collected in clinical settings are still limited in number. We have collected and analyzed a depression interview dataset labelled by clinicians, with spontaneous responses from 113 outpatients. This dataset provides a valuable and abundant resource for other researchers working on automated depression diagnosis, affective computing, and other related fields. It is especially useful for researchers who have difficulty accessing qualified psychiatrists to diagnose and label their interview recordings. Furthermore, this dataset can serve as a public benchmark for researchers who need to evaluate their models.

We trained baseline models for depression diagnosis and level classification on the dataset. The models achieved macro-average F1 scores of 0.85 and 0.81 for binary classification and 0.58 and 0.52 for depression level prediction using nested cross-validation. All models were evaluated by the nested cross-validation method and tested on an independent test set. These results demonstrate that automated depression diagnosis based on interviews in Chinese is feasible. Finally, we conducted statistical analysis with two different methods. Subjects were divided into groups based on depression severity, and intragroup feature analyses were completed. We confirmed acoustic features such as $F_0$, MCEP, and higher-order HMPDM significantly impact the ability to distinguish between depressed and healthy individuals. Moreover, a novel visualization method is presented to illustrate the high-impact audio features in depression detection, which further highlight the black-box nature of our proposed models as well as providing a reference for physicians. We anticipate the release of this dataset will motivate additional researchers to work on new models for automated depression diagnosis based on Chinese. We hope our dataset can also become a benchmark for other researchers to compare the performance of their models against others’ and supplement other datasets collected under controlled lab settings. We envision this dataset providing greater insight into AI for mental healthcare for mental health researchers and professionals.
Fig. 8: The contribution of audio features when making inferences about depressive individual S056, S057, S077, S080.

ACKNOWLEDGMENT

We would like to thank the funding support of MITACS for this research. This research is also supported by the China Scholarship Council (CSC) No. 202000810031.

REFERENCES

[10] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan,


Fig. 10: The contribution of audio features when making inferences about the healthy individual S084, S090, S091, S098.


Kaining Mao  Kaining is a Ph.D. student in the Electrical Computer Engineering Department at the University of Alberta. His work focuses specifically on Multimodal Machine Learning (MMML) and its impact on Automated Depression Diagnosis (ADD).

Tiansheng Zheng  Dr. Tiansheng Zheng has a Master’s degree in psychiatry and has been accredited as an Associate Chief Physician in psychiatry. She currently serves as the Deputy Director of Depression Center and Sleep Center of the Affiliated Kangning Hospital of Wenzhou Medical University. Her research interests include cognitive function in patients with depression and insomnia disorders.

Deborah Baofeng Wang  Dr. Deborah Baofeng Wang is responsible for the management of several international R&D projects at Wenzhou Kangning Hospital, focusing on the application of AI in mental health. With a Ph.D. in educational psychology and an MBA/management information system concentration, Dr. Wang has extensive experiences working as a Senior Research Analyst in the fields of mental health, public health, and education handling longitudinal and cross-sectional data at national (US) and international levels.

Lei Qian  Mr. Lei Qian holds a Master’s degree in clinical psychology and is accredited as an Intermediate Psychotherapist. He also serves as the Deputy Director of Higher Education Collaborations with the Affiliated Kangning Hospital of Wenzhou Medical University. His research interests include data analysis and mental health.

Rongqi Jiao  Rongqi Jiao is currently a Master’s student at Wenzhou Medical University. With a Bachelor’s degree in Medicine and the opportunity to learn as a psychiatrist at Wenzhou Kangning Hospital, Ms. Jiao has systematic academic training in medicine and hands-on research and clinical experiences in the area of mental health.

Wei Lyu  Accredited as a Chief Physician in psychiatry and Intermediate Psychotherapist, Dr. Wei Lyu currently serves as the Vice President of Wenzhou Kangning Hospital and an adjunct professor at the School of Mental Health, Wenzhou Medical University; He has extensive practical experiences as a psychiatric practitioner. He is also actively involved in the field of mental health as a professional as well as a leader. He is also a Youth Member with the Society of Psychiatrists, Zhejiang Medical Doctor Association.

Yanhui Zhu  Yanhui Zhu is a Master’s student majoring in psychiatry and mental health in the School of Mental Health, Wenzhou Medical University. The primary focus of her research in the Master’s program has been on microexpression and mental illness. She is currently working in the Second Affiliated Hospital of Wenzhou Medical University, as a clinical practitioner in mental health.

Jie Chen  Jie Chen received his Ph.D. degree from the University of Maryland at College Park, USA. He is currently a Professor at the University of Alberta. Dr. Chen is an IEEE Fellow, a Fellow of the Canadian Academy of Engineering, a Fellow of the American Institute of Medical and Biological Engineering, and a Fellow of the Asia-Pacific Artificial Intelligence Association. He received the distinguished Alumni Award, the University of Maryland. He was also awarded Killam Annual Professorship, which is among the highest honours given to a professor at Canadian Universities.

Bin Wu  Bin Wu is currently a Master’s student at Wenzhou Medical University. He holds a Bachelor’s degree in Medicine and is now also working as a psychiatrist at Wenzhou Kangning Hospital. He has received systematic academic training in the area of mental health and has solid hands-on experience in clinical practice of mental health, particularly the application of AI in mental health.

Minjie Ye  Dr. Minjie Ye is the Vice President and Director of the Center for Child and Adolescent Mental Health Services of the Affiliated Kangning Hospital of Wenzhou Medical University. He also serves as the Academic Lead for Kangning Hospital’s National Key Clinical Specialty (Psychiatry). In addition to his position as the Associate Dean of the School of Mental Health at Wenzhou Medical University, Dr. Ye also heads the university’s Child Mental Health Research Center. Dr. Ye is a member of the Society of Psychiatrists at Chinese Medical Doctor Association as well as a member of Committee of Marriage and Family Psychology with the Chinese Association of Social Psychology.