Advancing Automated Depression Diagnosis: Multimodal Analysis and a Novel Clinical Interview Corpus with Guidelines for Reproducibility and Generalizability

by

Kaining Mao

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Integrated Circuits and Systems

Department of Electrical and Computer Engineering
University of Alberta

Abstract

Depression is a major public health issue globally and is challenging to diagnose and treat in the early clinical stage due to the lack of understanding of the pathogenic mechanism. Traditional diagnosis heavily relies on physicians' experience and is subject to bias. With the advancement of smart devices and artificial intelligence, understanding how depression associates with daily behaviors can be beneficial for early-stage diagnosis and reduce the likelihood of clinical mistakes as well as physician bias. In this thesis, the author proposes an attention-based multimodality speech and text representation for depression prediction using the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) dataset.

First, the author conducted a review of studies from the past decade that utilized speech, text, and facial expression analysis to detect depression. The review includes information on the number of participants, techniques used to assess clinical outcomes, speech-eliciting tasks, machine learning algorithms, metrics, and other important discoveries for each study. A database has been created containing the query results and an overview of how different features are used to detect depression.

Furthermore, the author's model is trained to estimate the depression severity of participants using acoustic and semantic features. For the audio modality, the author uses the COVAREP features provided by the dataset and employs a Bi-LSTM followed by a Time-distributed CNN. For the text modality, the author uses GloVe to perform word embeddings and feeds the embeddings into the Bi-LSTM network. The results show that both audio and text models perform well on the depression

severity estimation task, with the best sequence level F_1 score of 0.9870 and patient-level F_1 score of 0.9074 for the audio model over five classes (healthy, mild, moderate, moderately severe, and severe), as well as sequence level F_1 score of 0.9709 and patient-level F_1 score of 0.9245 for the text model over five classes. Results are similar for the multimodality fused model, with the highest F_1 score of 0.9580 on the patient-level depression detection task over five classes.

In addition, the author presents a novel multimodal corpus comprising interviews conducted with clinically depressed patients, gathered directly from a psychiatric hospital. The dataset contains 113 interview recordings with 52 healthy and 61 depressed patients, and each data sample is annotated by experienced physicians, generating a binary label of depression versus healthy and a MADRS score. The author built baseline models to detect and predict depression presence and level, and the decision-making process of the model is investigated and illustrated.

In summary, the author conducts a review of studies utilizing speech, text, and facial expression analysis to detect depression and provides guidelines for collecting data and training machine learning models to ensure reproducibility and generalizability across different contexts. The author also proposes an attention-based multimodality representation, integrating speech and text modalities, for predicting depression. Additionally, they present a novel multimodal corpus of clinical interviews focused on depression. The author's work contributes to the advancement of automated depression diagnosis and treatment, which is critical in addressing the global public health issue of depression.

Preface

This dissertation is submitted for the degree of Doctor of Philosophy at the University of Alberta. This Ph.D. thesis is based on the research performed at the Department of Electrical and Computer Engineering, University of Alberta, under the supervision of Professor Jie Chen from September 2019 to April 2023.

Chapter 2 has been published as "A survey on Automated Clinical Depression Diagnosis" in npj Mental Health Research Res 2, 20 (2023), doi: 10.1038/s44184-023-00040-z, with Kaining Mao, Yuqi Wu, Jie Chen as authors. I was responsible for literature review, data analysis, and manuscript composition, while Dr. Jie Chen provided valuable guidance and revisions.

Chapter 3 contains the research published as "Prediction of Depression Severity Based on the Prosodic and Semantic Features With Bidirectional LSTM and Time Distributed CNN," in IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 2251-2265, 1 July-Sept. 2023, doi: 10.1109/TAFFC.2022.3154332., authored by Kaining Mao, Wei Zhang, Deborah Baofeng Wang, Ang Li, Rongqi Jiao, Yanhui Zhu, Bin Wu, Tiansheng Zheng, Lei Qian, Wei Lyu, Minjie Ye, and Jie Chen. I designed the experiments, analyzed the data, and wrote the manuscript, with Dr. Jie Chen supervising the work, provided valuable guidance and revisions.

Chapter 4 has been published as "Analysis of Automated Clinical Depression Diagnosis in a Chinese Corpus," in IEEE Transactions on Biomedical Circuits and Systems, vol. 17, no. 5, pp. 1135-1152, Oct. 2023, doi: 10.1109/TBCAS.2023.3291554., with Kaining Mao, Deborah B. Wang, Tiansheng Zheng, Rongqi Jiao, Yanhui Zhu,

Bin Wu, Lei Qian, Wei Lyu, Jie Chen, and Minjie Ye as authors. For this chapter, I designed and conducted the experiments, analyzed the data, and wrote the manuscript, with Dr. Jie Chen supervising the work and providing valuable guidance and revisions. The study was approved by the ethics committee of Wenzhou Kangning Hospital (No. AF/SQ-02/01.0).

Acknowledgements

I am deeply grateful to my supervisors, Dr. Jie Chen and Dr. Xihua Wang, whose strong support and insightful guidance have been essential throughout my academic journey. Their dedication to mentoring and their invaluable feedback have played a pivotal role in shaping my thesis and ultimately earning my degree.

I am also thankful to the members of my supervisory committee, Dr. Chintha Tellambura, Dr. Cor-Paul Bezemer, Dr. Xingyu Li, Dr. Yanbo Zhang and Dr. Xiaoping Zhang, for their constructive critiques and suggestions, which have helped refine my work to its current form.

My appreciation extends to my colleagues in the research group: Jiajun Li, Shuren Wang, Shiang Qi, Wei Zhang, Xuanjie Ye, Yuqi Wu, Zuyuan Tian, Zixuan Wang, Tianxiang Jiang, Zuyuan Tian, Bingxuan Li, Yiwei Feng, and Steven Bai. Collaborating with such diligent and talented researchers is my pleasure.

Special thanks go to my partner, Yuan You, for her support and kindness during challenging times. Together, we have explored many breathtaking destinations, and our journey continues to lead us to more charming places.

Additionally, I am deeply thankful to my parents and family for their support and encouragement, without which my academic pursuits would not have been possible.

I am also grateful to the China Scholarship Council and the University of Alberta for their financial assistance, as well as to Wenzhou Kangning Hospital for their support during data collection, clinical trials, and patent applications.

My appreciation goes to all those who have helped me in my doctoral studies,

including the anonymous reviewers whose important feedback improved my papers. Their contributions are invaluable in shaping my academic journey.

Lastly, I owe a special thanks to my past self from four years ago, for choosing to pursue a PhD degree and persevering through failures and challenges, which ultimately led to my achievements. Without that initial determination, I would not be where I am today.

Table of Contents

List of Tables

List of Figures

Abbreviations

ADD Automated depression diagnosis.

AI Artificial intelligence.

API Application programming interface.

ASR Automatic speech recognition.

AVEC The Audio/Visual Emotion Challenge.

BDI Beck's Depression Inventory.

Bi-LSTM Bidirectional Long Short-Term Memory.

CCC Concordance correlation coefficient.

COVAREP Cooperative Voice Analysis Repository.

DAIC-WOZ Distress Analysis Interview Corpus.

DSM Diagnostic and Statistical Manual.

EEG Electroencephalogram.

HMPDM Harmonic-to-Noise Ratio Partial Distance Measure.

LSTM Long Short-Term Memory.

MADRS Montgomery-Asberg Depression Rating Scale.

MCEP Mel-cepstral coefficients.

MDQ Maxima Dispersion Quotient.

MSE Mean square error.

NAQ Normalized Amplitude Quotient.

OQ Open Quotient.

PHQ Patient Health Questionnaire.

PRISMA Preferred Reporting Items for Systematic reviews and Meta-Analyses.

PSP Parabolic Spectral Parameter.

PTSD Post-traumatic stress disorder.

QOQ Quasi Open Quotient.

RMSE Root mean square error.

ROC Receiver operating characteristic curve.

SHAP SHapley Additive exPlanations.

TD-CNN Time-distributed convolution neural network.

Uni-LSTM Unidirectional Long Short-Term Memory.

VUV Voiced/Unvoiced.

WHO World Health Organization.

Chapter 1

Introduction

In recent years, the combination of information technology and mental health research has opened new avenues for understanding and addressing the challenges associated with mental health disorders. Among these disorders, depression stands out as a pervasive and debilitating condition that affects millions worldwide. Our study investigates into the integration of acoustic and semantic features extracted from depression interviews, aiming to develop a deep learning model for assessing depression levels. By leveraging deep learning models, we seek to explore the intricate relationship between speech patterns, linguistic content, and the depressive symptoms. This thesis investigates the potential of multimodality approach, bridging the gap between traditional diagnostic methods and cutting-edge technology, to enhance our ability to identify and understand depression on a deeper level. Through the fusion of acoustic and semantic features, our research endeavors to contribute valuable insights to the research of mental health assessment, with the ultimate goal of facilitating self-assessment for depression risk and prompt users to seek appropriate mental healthcare.

1.1 Application of Artificial Intelligence on Depression Screening

1.1.1 Automated Depression Diagnosis

Automated depression diagnosis presents a promising solution to overcome the barriers hindering timely diagnoses and treatment for mental health issues. Mohr et al. reported that stigma, lack of motivation, time or availability constraints, and cost as the main obstacles faced by patients seeking help [1–3]. Recent research by Schuller et al. has shown that high-speed networks and smartphones with high-performance computational units can support continuous monitoring of the psycho-emotional state over a prolonged period [4, 5]. Various models have been developed to detect depression and other psychological disorders by extracting features from interview videos, audio recordings, neuroimaging data, social media posts, and transcribed audio recordings [6–18]. Toolkits such as OpenFace can extract facial landmarks, action units, face orientation, and eye gaze [19]. Other modalities, such as neuroimaging data, have been used to predict the presence of Schizophrenia [20]. Lastly, features extracted from social media and transcribed audio recordings have also been used to detect depression and stress [21, 22]. These technological advancements highlight the growing need for affordable screening techniques. Social media platforms offer an opportunity to leverage automated systems for identifying potential patients. Automated diagnostic tools enable individuals with depression who have not sought professional help to remotely evaluate their mental states and receive online support from healthcare workers. These tools can also be designed to customize treatment based on an individual's specific symptoms, thus improving treatment efficacy [23, 24]. Furthermore, such systems can be employed for mental disorder screening in various settings, including universities, the military, and basic healthcare facilities.

1.1.2 Benefits of Automated Depression Assessment

Artificial intelligence (AI) based depression detection systems have been extensively studied, focusing on data collection, emotion induction, and prediction of depression using multiple modalities. The potential of AI to revolutionize the diagnosis and prognosis of mental health disorders is significant. By utilizing large and diverse datasets, AI models can be trained to accurately screen for early-stage mental illnesses, thereby providing a crucial tool for enhancing the overall mental health of the population. The benefits of automated depression assessment tools are supporting clinicians in making accurate diagnoses and providing effective treatment, identifying at-risk individuals before they seek treatment, and tracking symptoms over time, both during and after treatment. Automated depression assessment systems also aid doctors in diagnosing depression and making related decisions. The Research Domain Criteria, created by the National Institute of Mental Health, assists in distinguishing diagnoses and symptoms [25]. Consequently, we can train models to predict the probability of different mental disorders to assist clinicians in diagnosing and early intervention if suicidal thoughts are detected. Lastly, automated depression assessment models facilitate mental healthcare by enabling more frequent and real-time symptom monitoring. Real-time monitoring ensures that individuals at risk for depression are reminded to seek mental healthcare and allows for the detection of important signs related to suicidal or self-harm thoughts. Online psychotherapy can also be conducted based on real-time monitoring, enabling timely interventions and tailored treatment plans. Furthermore, these models enable customized treatment plans based on multimodality features (genetic, behavioural, neuroimaging) [26–28]. Previous articles have demonstrated that multimodality models usually outperform unimodality models [29–32]. As a result, automated depression assessment models improve the efficiency of the healthcare systems, lower costs, and make treatment plans more customizable.

1.1.3 Challenges and Limitations of Automated Depression Assessment

While there are numerous potential benefits to automated depression assessment, certain challenges need to be addressed to fully realize its advantages. Previous studies have relied on small and non-representative 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 datasets have investigated music-induced [33], video-induced [29, 34] and mixed emotion induction methods [35]. 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. Additionally, models can be biased and lack interpretability, which limits their clinical applicability. For example, the model may tend to assign lower depression likelihood in male subjects because fewer male subjects have depression in the training set [36–40]. To address these issues, there is a need for interactive multimodal datasets for the study of depression, collected through interviews conducted in a clinical setting between patients and physicians. Moreover, a new architecture is essential to improve the generalizability of previous models.

1.2 Understanding Depression: Prevalence, Impact, and Diagnostic Challenges

1.2.1 Prevalence and Impact of Depression

The prevalence of mental health disorders, such as depression, and the associated challenges in diagnosis and treatment have drawn increased attention in recent years. The COVID-19 pandemic has further exacerbated the prevalence of depression and anxiety among the general population. According to the World Health Organization

(WHO), clinical depression is predicted to become the second most debilitating disease by 2030, ranking only behind cardiovascular diseases [41]. It is estimated that the average cost of treating depression in 2010 is € 24, 000 per patient and the total cost can be as high as € 92 billion in Europe [42]. In the United States, depression causes an estimated loss of $44 billion, due to absence or low working efficiency [43]. Suicide is one of the severe results of depression, and the WHO reports that the number of people who passed away due to suicide is over 800, 000 every year [44]. The attempted suicide is more frequent, possibly no less than 20 times that of those who died by suicide [44]. Patients with depression are more apt to generate suicide thoughts [45, 46]. It is estimated that more than 50% of people who died by suicide meet clinical criteria of depression [47, 48].

1.2.2 Challenges in Diagnosis and Treatment

Unfortunately, the symptoms of depression are not always apparent. Many individuals who do not have depression may exhibit sadness and hopelessness, while those who do suffer from depression often hesitate to report their condition and seek treatment. In 2017, the WHO reported that over 264 million people of all ages were affected by depression, with 75% of individuals in low and mid-income countries unable to receive qualified psychotherapy [49]. Stigma resulted by depression leads to individuals with depression hiding their symptoms. On the other hand, depression diagnosis is challenging due to episodic symptoms and multiple co-occurring disorders, as demonstrated by the low inter-rater reliability [50] and test-retest reliability scores [51] in major depressive disorder diagnosis. A delayed or inaccurate diagnosis can have severe consequences, including the potential for suicide. Moreover, traditional methods for assessing and monitoring depression involve subjective, semi-structured interviews between patients and healthcare professionals, which can be influenced by bias, cognitive limitations, and social stigma. Additionally, economic conditions and living

constraints often prevent depressed individuals from accessing qualified psychological treatment.

Therefore, the development of low-cost screening techniques that can be deployed in communities and operated by non-specialists would be highly beneficial. Early-stage detection of mental disorders is crucial for individuals, policymakers, and security agencies due to the association of these disorders with adverse behaviors, including mass shootings [52].

1.3 Summary

In conclusion, the prevalence of mental health disorders, particularly depression, calls for innovative solutions to overcome the barriers to timely diagnosis and treatment. Automated depression assessment is promising in addressing these challenges, benefiting individuals, healthcare professionals, and the healthcare system. By leveraging technology advancements and diverse datasets, AI-based models can play a pivotal role in accurate screening, personalized treatment, and real-time monitoring of depression. However, further research and development are necessary to address the limitations and challenges associated with automated depression assessment systems. In this thesis, we firstly reviewed recent research on using machine learning methods using acoustic, semantic, and facial features. Reviews have been written on predicting depression or suicidal risk using speech cues; however, our review stands out for its use of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines for an extensive and rigorous evaluation of the latest research findings. Other than the review, we proposed a multimodality automated depression diagnosis system with prosodic and semantic features to predict the depression levels. Unlike previous models, our proposed model does not have a strict limitation of input duration. This mitigates the problem that the audio/text feature sequences are required to be the same in length in previous articles. Moreover, we addressed

the challenges and limitations of existing depression interview dataset. For instance, most previous depression interview datasets are non-clinical, labelled with self-report depression rating scale, recorded in controlled conditions, and produced in English. In comparing our dataset to others, our approach collects spontaneous responses from participants in clinical setting. Our dataset can also be a valuable resource for investigating the impact of culture difference on automated depression detection models.

1.4 Contribution and Novelty of This Thesis

We initially reviewed the current research on AI for depression screening. This review is unique in that it uses the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines for extensive and rigorous evaluation of the latest research findings. By synthesizing and analyzing recently published data, this review offers important insights into the current state of the field and identifies key areas for future research.

Our first original research proposes an automated depression diagnosis system that uses prosodic and semantic features to predict depression levels. We combine Bi-LSTM and TD-CNN models to achieve this. This is the first time that a time-distributed CNN has been used to extract temporal information from the output of an LSTM encoder. Our model can predict depression levels for a patient-independent audio or text feature sequence of any length, as long as the number of features meets our specification. Our system provides a series of estimations of depression severity based on the audio/text feature, which are merged through a major voting algorithm to produce a patient-level depression severity prediction. Unlike previous articles, our model does not require audio/text feature sequences of the same length. We use the bidirectional LSTM model to learn long-term bidirectional dependencies in audio and text feature sequences, and the time-distributed CNN architecture to learn the spatial features of data. By combining the strengths of both models, our hybrid

LSTM and TD-CNN model performs well in learning the spatiotemporal sequence. Our audio and text models achieved patient-independent F_1 scores of 0.9870 and 0.9709, respectively, on the test partition of the DAIC-WOZ dataset. The fused multimodality model achieved the best F_1 score of 0.9580 on the same dataset.

Our second research focuses on investigating the effectiveness of semantic and prosodic features in evaluating depression risk in other languages. To ensure a high-quality dataset, we conducted interviews in Chinese between clinicians and outpatients, and evaluated patients using the Montgomery-Asberg Depression Rating Scale (MADRS) [53]. We extracted audio features, such as formant frequency F_0 and Normalized Amplitude Quotient (NAQ), using the COVAREP toolbox, and transcribed the interview recordings using an audio transcription application programming interface (API) developed by iFlyTek [54]. Our dataset includes both the interview recordings and their transcripts, making it the first multimodal clinical distress interview corpus with over 100 subjects in Chinese. Our analysis shows that there is a significant difference in 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 demonstrate that a subset of acoustic features has strong discriminative ability in 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. We calculate the influence of each acoustic feature and list them in descending order, providing new insights for physicians to focus on distinguishing depression severity among patients.

1.5 Thesis Outline

This thesis is organized into five chapters. Chapter 1 provides an overview of the importance of automated depression diagnosis and presents the research question and

objectives of this thesis. Additionally, this chapter highlights the contributions and novelty of this work. Chapter 2 reviews the progress made in the field of automated depression diagnosis, including the challenges and limitations of current approaches. This chapter sets the stage for the subsequent chapters and provides context for the research presented in this thesis. Chapter 3 focuses on depression detection and assessment based on a public depression interview dataset. This chapter introduces the dataset and presents our proposed multimodality fusion model for predicting the presence and severity of depression. Chapter 4 presents the Chinese depression interview corpus, the largest clinical depression interview dataset to date. We analyzed this dataset and provide baseline results, which can serve as a benchmark for future research. Finally, Chapter 5 summarizes the research presented in this thesis and discusses possible future work in the field of automated depression diagnosis. Through this work, we hope to contribute to the development of more effective and accessible methods for diagnosing depression.

Chapter 2

A Systematic Review on Automated Clinical Depression Diagnosis

In this chapter, we conducted a thorough review of studies from the past decade that focused on using speech, text, and facial expression analysis to detect depression, as defined by the Diagnostic and Statistical Manual (DSM-5). Following the PRISMA guideline, we provide key details for each study, including participant numbers, clinical assessment techniques, speech tasks, machine learning methods, metrics, and other significant findings. Following the PRISMA guideline, we provide key details for each study, including participant numbers, clinical assessment techniques, speech tasks, machine learning methods, metrics, and other significant findings. We have compiled this information into a database for easy reference, summarizing how different features are used to detect depression. Given the diversity in datasets, feature extraction methods, and metrics in this field, we have outlined guidelines to ensure reproducibility and generalizability when collecting data and training machine learning models across different situations. This chapter has been published as "A survey on Automated Clinical Depression Diagnosis" in npj Mental Health Research Res 2, 20 (2023), doi: 10.1038/s44184-023-00040-z.

2.1 Inclusion Criteria and Literature Search

The PRISMA guidelines were followed in this literature review, as shown in Figure 2.1. Our goal was to search for articles published in the last ten years that included artificial intelligence methods for predicting the presence or severity of major depressive disorder by analyzing acoustic, semantic and facial landmarks. Google Scholar was used as the search engine for articles from 2012 to the present, queried between July 20, 2022, and May 20, 2023, excluding case studies, studies that solely used perceptual evaluation of speech, studies without a control group or clinical depression rating scales, non-peer-reviewed preprint and theses, and articles published before 2022 and having fewer citations than years of publication (e.g. articles published in 2019 with three citations, or articles published in 2017 with five citations would be included). We excluded certain articles that lacked comprehensive methodology or detailed results. Additionally, we encountered cases where articles were published in both journals and conference proceedings, covering similar topics, methods, and results. Furthermore, some articles only focused on proposing methods for feature extraction without incorporating the training of models for depression detection. The search terms used to find relevant articles were: "allintitle:(("depression" OR "major depressive disorder") + (acoustic OR acoustical OR speech OR voice OR vocal OR audio OR pitch OR prosody OR prosodic OR vowel) + (automated OR behavioural OR measures OR diagnosis))." Articles related to depression caused by Parkinson's Disease, autism, and substance overdose disorders were excluded. Replacement of the acoustic feature with the semantic feature and facial landmarks in the command resulted in the following search term with associated features: "allintitle:(("depression" OR "major depressive disorder") + (semantic OR text OR interview OR transcript OR social media) + (automated OR behavioural OR measures OR diagnosis))" and "allintitle:(("depression" OR "major depressive disorder") + (facial expression OR visual OR facial features OR facial landmarks OR facial muscles)+ (automated OR

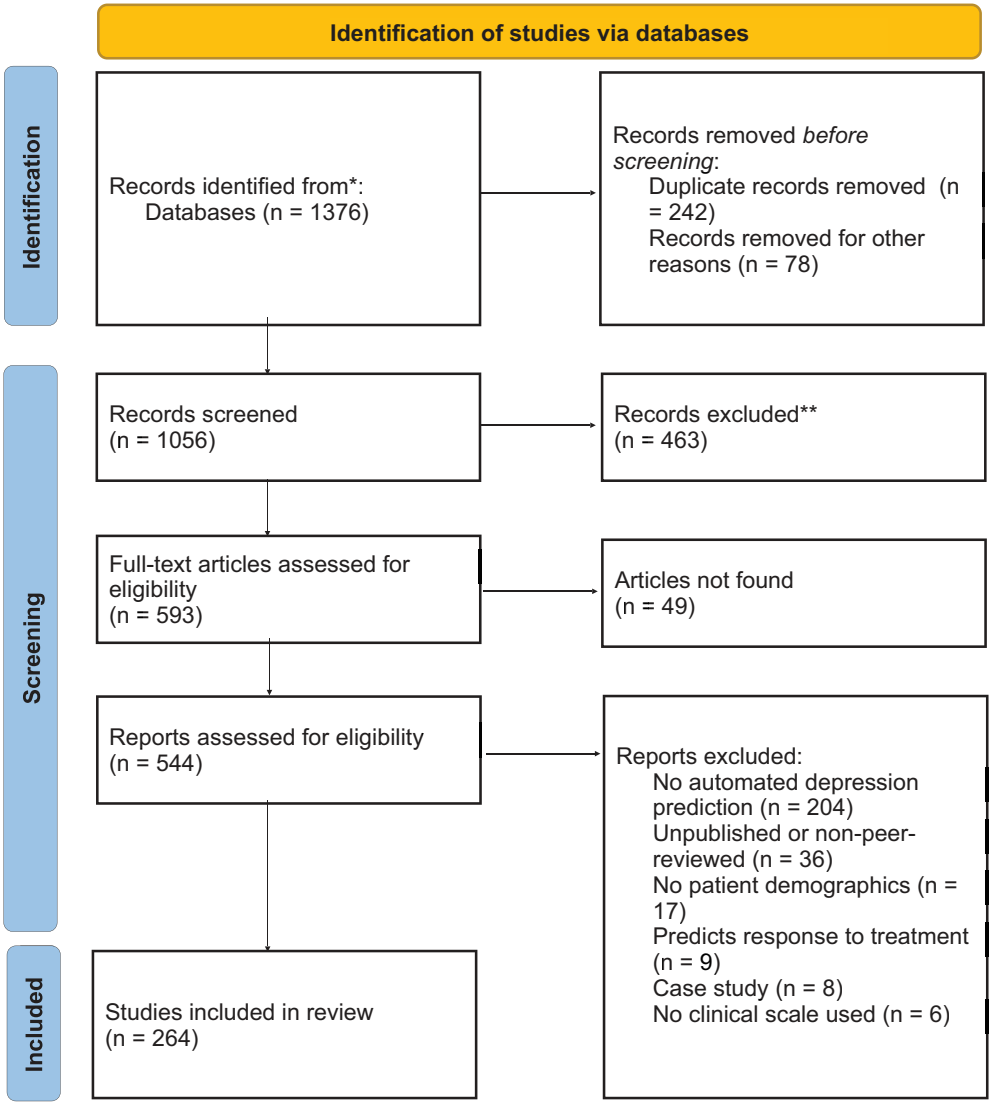

Figure 2.1: PRISMA flow diagram of study inclusion and exclusion criteria in this review.

Table 2.1: Summary of Literature Review Results

Modality	Articles	Median dataset size (range)	Clinical assessment	Predictive models
Acoustic	140	189	36	140
Semantic	99	1046	2	81
Facial landmarks	25	49	16	21

behavioural OR measures OR diagnosis))."

Information extraction was performed by reading the title, abstract and conclusion. The following information was synthesized from each article: mental disorders, number of subjects, age range, optimal model, best metrics, type of validation and predictive features. Due to the limited number of studies we were able to review and include in this review, we only searched for keywords in the titles of articles rather than in other sections. The screening process of the articles involved reading the title and abstract. Only articles that were relevant to using machine learning to detect depression and had "machine learning" or related terms in the title were included, and the others were excluded. Our study may not have captured all relevant articles on this topic, and other studies not focused specifically on automated depression diagnosis using machine learning methods may have been missed.

2.2 Results

264 studies were included in the review. Table 2.1 summarizes the search results. Synthesized information can be found online https://bit.ly/3DBQtZk, https://bit.ly/43Q6Yvy and https://bit.ly/44IKaPv, which can be extended by adding new studies on a blank row or fields on a blank column. Previous review and datasets-only articles were included in this study but were not included in Table 2.1.

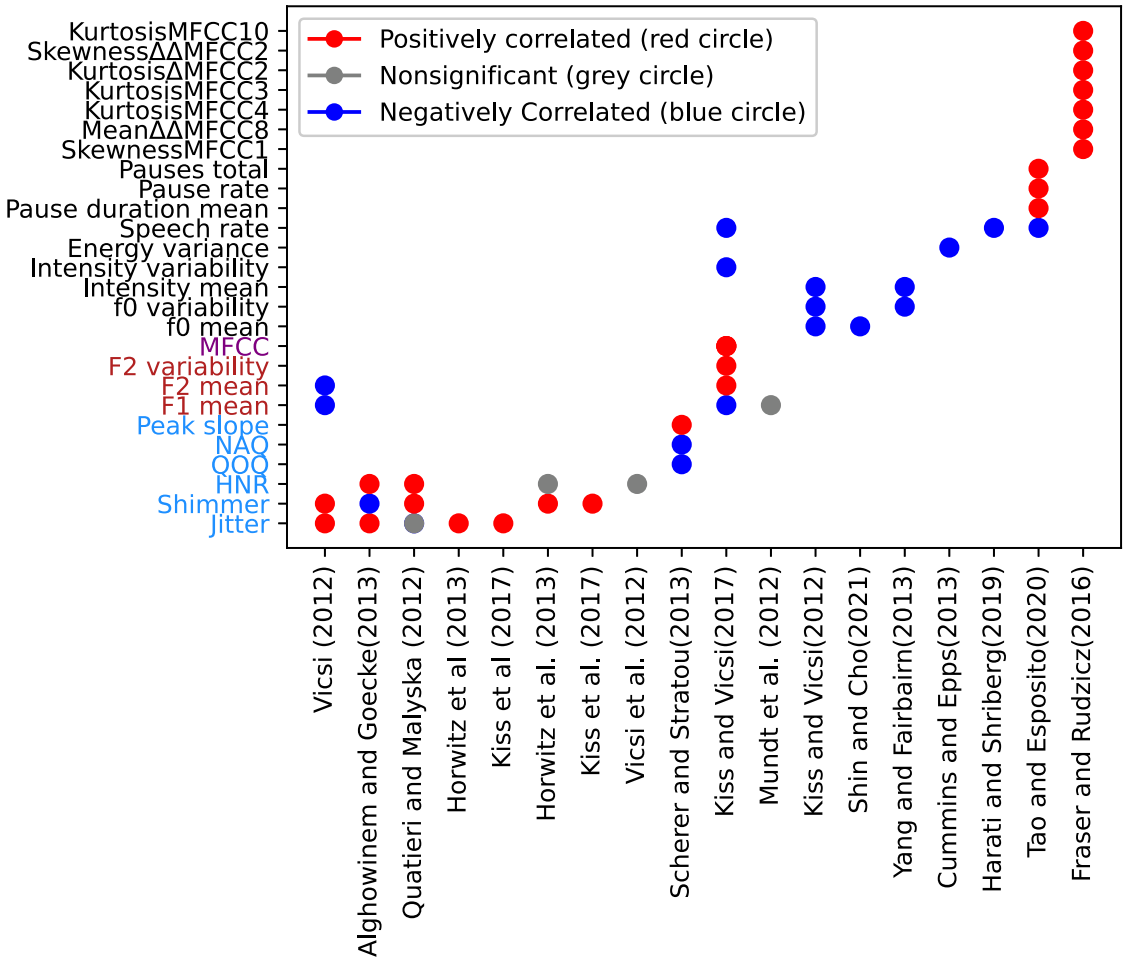

Figure 2.2: Synthesis of acoustic feature analysis in major depressive disorder. Acoustic features are sorted, such as vocal fold source features (blue), vocal tract filter features (red), spectral features (purple), and features related to prosody or melody (black). Features that are significantly higher in a psychiatric group than healthy controls or that correlate positively with the depression level receive a score of 1 (red), features that are lower or correlate negatively receive a score of -1 (blue), and nonsignificant or contradicting findings receive a score of 0 (gray). Features not studied in any studies are blank.

2.2.1 Predictive Acoustic Features in Major Depressive Disorder

The study of speech patterns has been a research topic in identifying indicators of mental disorders since the 1920s. Emil Kraepelin, the founder of modern scientific psychiatry, reported that the voices of depressed patients were lower in pitch, sound intensity, and speech rate, and instead tend to be monotonous and hesitant, with shuttering and whispering [55]. Unlike other behavioural features like EEG and ECG, speech expresses real emotion and thought more directly, making it harder for patients to hide symptoms. Moreover, acoustic features can be extracted across different languages, which is important for languages without pre-trained natural language processing models. In addition, speech recordings can be easily collected with smartphones and laptop computers instead of complex and costly equipment. With the advancements in speech recognition, especially its application for electronic medical records, speech recording will become more accessible for research purposes. With the publicly available code provided by Low et al. [56], we created Figure 2.2, which provides a synthesis of the acoustic features investigated using machine learning. The table shows the acoustic features found to be statistically different between the group with a mental disorder and the healthy control group or highly correlated with a diagnostic rating scale. Each cell in Figure 2.2 represents the correlation between a specific acoustic feature and depression. For example, an acoustic feature that correlates positively with the disorder severity would be marked with a red dot, a negative correlation with a blue dot, and a non-significant feature with a grey dot.

Table 2.2 provides an overview of the key findings from previous studies on automated depression detection using acoustic features. One common finding among the studies is the relationship between acoustic volume and depression. Cummins et al. [57–59] found that as the level of depression increases, the acoustic volume significantly decreases, indicating a potential acoustic marker for depression.

Another notable finding is the influence of gender (biological sex) on acoustic features related to depression. Cummins, Morales and Vicsi et al [60–62] proposed gender (biological sex)-dependent formant features that outperformed acoustic-only features in depression detection. This suggests that gender-specific acoustic characteristics may play a role in accurately detecting depression. Kiss et al. [63] highlighted the importance of speech rate, articulation rate, pause lengths, and formant frequency in detecting depression. They found that these acoustic features differed between individuals with depression and the control group, suggesting potential utility in automated depression detection. Stasak et al. [64] proposed that depressed individuals tend to use phonemes that require less effort and demonstrate decreased articulatory precision. These findings indicate that analyzing articulatory characteristics could provide valuable insights for depression detection.

It is important to note that these findings are based on previous studies and should be interpreted within the context of their respective methodologies and limitations. Further research is needed to validate and refine the use of these acoustic features for automated depression detection.

2.2.2 Predictive Semantic Features in Major Depressive Disorder

Conventionally, depression is detected using clinical depression rating scales administered by clinicians. However, these rating scales have limitations, as responses can be influenced by factors such as the patient's emotional state, relationship with the clinician, and patient self-bias (e.g. participants may be more likely to exaggerate their symptoms) [67]. With the advancement of machine learning applied to text data from social media, new methods have emerged to address these limitations. Social media such as Twitter, Facebook and Reddit provide a wealth of information about individuals' feelings, thoughts and activities. Machine learning, especially text mining and sentiment analysis techniques, have become more accurate and intelligent, aid-

Table 2.2: Predictive Acoustic Features in Prior Research Publications

Study	Main Findings
[57]	Decreased acoustic volume More concentrated MFCC space
[60]	Gender (biological sex)-dependent formant features
[61]	Fundamental frequency Pronoun use and negatively-valenced words
[65]	Tenser voice
[62]	Jitter and shimmer values of vowels First and second formant frequencies
[66]	Seeking care for voice problem
[58]	Acoustic volume Probabilistic acoustic volume slope
[59]	Lower voices Variance in voice pitch
[63]	Articulation rate Speech rate Pause lengths Formant frequency
[64]	Use phonemes that require less effort Articulatory precision

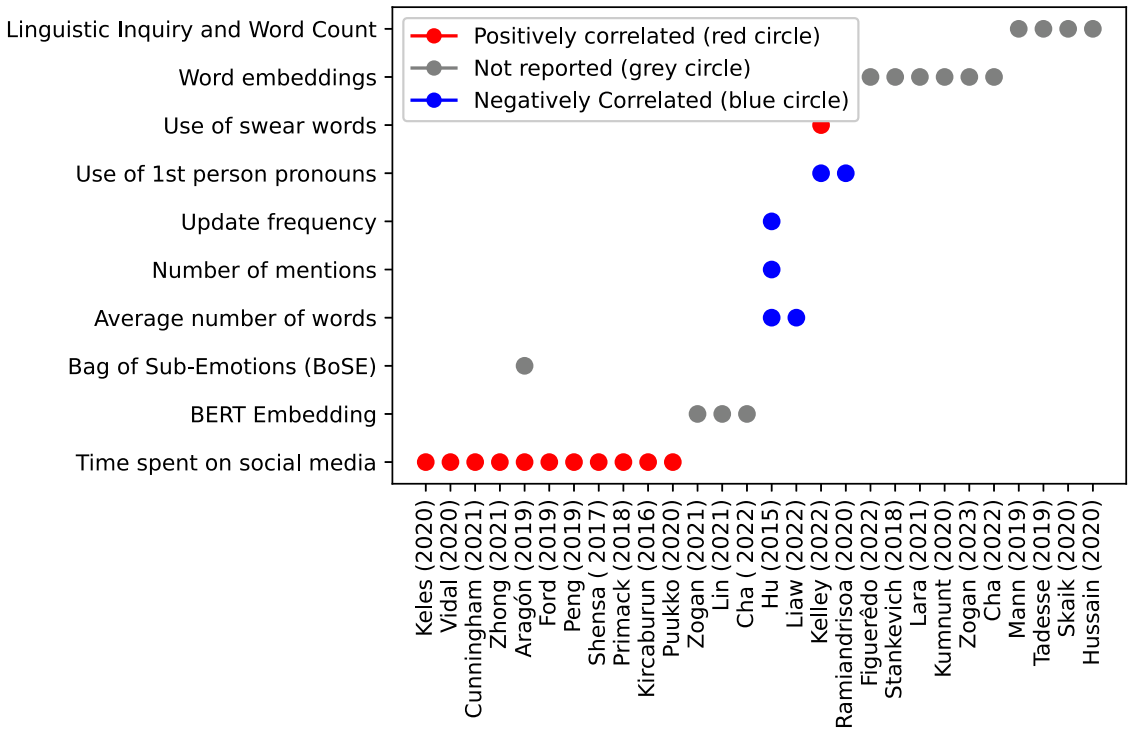

Figure 2.3: Synthesis of semantic feature analysis in major depressive disorder. Features that are significantly higher in a psychiatric group than healthy controls or that correlate positively with the depression level receive a score of 1 (red), features that are lower or correlate negatively receive a score of -1 (blue), and findings without reporting their changes receive a score of 0 (gray). Features not studied in any studies are blank.

ing mental healthcare providers to detect depression [68–74]. This summary reviews studies on automated depression detection using language cues. Previous studies identified depression based on clinician diagnosis, patient self-reported mental status and online forum memberships. Clinician diagnosis means that depression levels were determined by a clinician based on interview transcripts or online posts. Among the 99 studies using language cues, only two identified depression based on the clinician diagnosis [75, 76], while 77 studies used self-reported depression rating scales. The remaining 20 studies did not report which criterion were used to determine depression levels.

Social Media and Depression

Table 2.3 provides a summary of key findings from various studies examining the relationship between social media usage and depression. These studies employ a range of techniques and models to improve the detection and understanding of depression based on social media data. Several studies highlight increased social media usage among individuals with depression. Various advanced models, such as GRU models with knowledge-aware dot-product attention [77] and DeepBoSE [78], demonstrate improved performance in depression detection compared to conventional methods. Additionally, semantic mapping of emoticons [79] and the application of semantic role labeling [80] are proposed as techniques to enhance detection accuracy. These findings highlight the potential of leveraging machine learning and natural language processing techniques to gain insights into mental health conditions through social media data.

Furthermore, the studies presented in the table emphasize the importance of considering multimodal data, user characteristics, and sentiment analysis for a comprehensive understanding of depression [81, 82]. They also propose the use of lexicon features and emotional information capture to improve depression detection [83, 84]. The fusion of lexical features and the development of bipolar feature vectors demon-

strate promising results in enhancing prediction effectiveness [85, 86]. Additionally, the studies suggest the potential of analyzing social signals and user timelines to capture semantic features for depression detection [87, 88].

By synthesizing these findings, we have gained a deeper understanding of the potential of social media data in detecting and understanding depression. These insights can inform the development of effective mental health interventions, improve clinical practice, and contribute to the responsible and ethical usage of AI in this domain. Overall, the findings contribute to our understanding of the complex relationship between social media use and depression, providing valuable insights for mental health promotion and clinical practice.

Machine Learning Models for Depression Detection

Automated depression detection algorithms have been a subject of study by various researchers. Table 2.4 provides a summary of these studies, highlighting machine learning models and performance metrics employed. Salas et al. conducted a comprehensive review of previous studies that utilized language cues and found that word embedding was the most commonly used linguistic feature extraction method, while the support vector machine was the most prominent machine learning model [113]. [75, 76, 114, 115] focused on using machine learning methods to detect depressive symptoms in social media data, highlighting their potential as complementary tools in public mental health practice.

However, some weaknesses and variations in the literature have been raised. McCrae et al. conducted a review of studies examining the relationship between social media use and depression symptoms, highlighting the need for comparative analysis due to variations in methods, sample sizes, and results across studies [116]. They also suggested that future research should incorporate longitudinal analysis, as most studies were cross-sectional. Heffer et al. found no predictive association between social media use and depressive symptoms over time, challenging the assumption that

Table 2.3: Exploring the Predictive Relationship Between Social Media Usage and Depression

Study	Main findings
[89–103]	Social media usage increases among depressive individuals
[79]	Semantic mapping of emoticons improves the performance.
[80]	Future work needs to involve applying semantic role labelling to obtain better results.
[78]	DeepBoSE outperforms conventional Bag-of-Features(BoF) representations.
[83, 104]	Proposed depression lexicons that distinguish depressive individuals.
[87]	Analyzing users' social signals could be considered for further analysis.
[105]	Topic modeling features such as liked tweets can be useful.
[85]	Fused the lexical features using a correlation-based metric to enhance prediction effectiveness.
[84]	Capture deep emotional information from the input embeddings with a pre-trained TextCNN.
[88]	The model captures semantic features from user timelines for depression detection.
[81, 106]	User characteristics and sentiment analysis improved depression detection performance.
[107–111]	Depressed users exhibit reduced online activities, increased negative sentiment, and self-focused pronoun usage.
[112]	Depressed individuals are more likely to compare themselves to others and dislike being tagged in self-perceived unflattering pictures.

social media use leads to depressive symptoms [117]. These contrasting perspectives call for further investigation and highlight the complexity of the relationship between social media use and depression.

In conclusion, while automated depression detection algorithms show promise, the field still faces challenges in terms of standardization, methodological variations, and the need for longitudinal analysis. Future research should address these limitations, conduct a comparative analysis, and explore the intricate mechanisms underlying the relationship between social media use and depression. Additionally, ethical considerations and the potential impact of using social media data for mental health assessment should be carefully examined. Advancements in this field can contribute to the development of effective and reliable tools for early detection and intervention in depression.

2.2.3 Predictive Facial Features in Major Depressive Disorder

Table 2.5 provides a summary of previous studies on automated depression detection using facial features. The studies examined various aspects of facial expressions and their relationship to depression. Among the 18 studies that utilized facial landmark features, 13 studies identified depression based on clinician diagnosis, while four studies used self-report depression rating scales. One study did not report the specific criterion used to determine depression levels.

In [127–131], researchers proposed new architectures, analyzed facial expressions in videos, and demonstrated the effectiveness of facial analysis for automated depression diagnosis, achieving an F1 score of over 80%. Hunter et al. [132] evaluated the eye-tracking patterns of individuals with non-clinical depressive symptomatology in processing emotional expressions, revealing distinct differences compared to healthy individuals.

In addition to the studies mentioned in the original paragraph, several more recent

Table 2.4: Analyzing the Efficacy of Machine Learning Models in Detecting Depression through Social Media Data

Study	Main findings
[75]	Achieved 68% accuracy and 72% precision in identifying clinical depressive symptoms using a semi-supervised statistical model.
[76]	Proposed a new computational model and achieved a recall of 0.904, precision of 0.909, and F1 score of 0.912.
[118–120]	Demonstrated that a multi-kernel support vector machine is the most appropriate approach to identifying depression in individuals using social media.
[121]	Latent semantic analysis shows a significant difference in writing topics depending on users' mental health.
[122]	A word2vec pre-trained word embedding and random forest classifier achieved their best performance with a 0.877 F_1 score.
[123]	Fusion model can detect moderate depression or higher with 0.92 recall and 0.69 precision.
[124]	Proposed a system to effectively detect depression using social media content with an accuracy of 88% and F_1 score of 93%.
[104]	Application accurately identifies indicators of depression in Facebook users with 94% accuracy.
[72]	Achieved 91% accuracy and F_1 score of 93% with a multi-layer perceptron algorithm and combined features.
[125]	Achieved an accuracy, recall, and precision of 91.7% using a combination of text-based features and machine learning techniques.
[126]	Facebook behaviors can be used to predict depression levels with an accuracy of 85% and F_1 score of 88.9%.

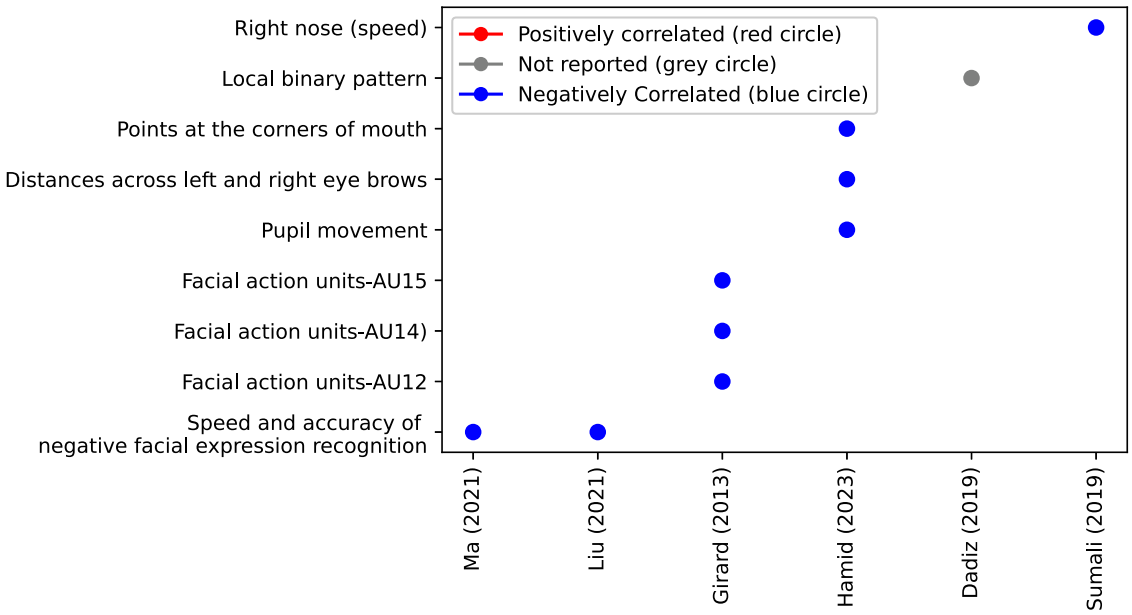

Figure 2.4: Synthesis of visual feature analysis in major depressive disorder. Features that are significantly higher in a psychiatric group than healthy controls or that correlate positively with the depression level receive a score of 1 (red), features that are lower or correlate negatively receive a score of -1 (blue), and findings without reporting their changes receive a score of 0 (gray). Features not studied in any studies are blank.

studies provide valuable insights into automated depression detection using facial features. Hamid et al. [133] designed a hybrid model that integrates electroencephalogram (EEG) data and facial features, surpassing existing diagnosis systems. Shangguan et al. [134] demonstrated that video stimuli and an aggregation method can be effective for automatic depression detection.

Overall, the summarized studies highlight the significance of facial expressions in automated depression detection. They showcase various approaches, including deep learning models, multimodal techniques, and analysis of specific facial features. These findings contribute to the understanding of how facial expressions can serve as valuable indicators for detecting and diagnosing depression.

2.3 Discussions

In this part, we're not just reviewing, but also suggesting ways to make previous studies better. Most studies in this literature review adopted automated speech feature extraction to assess major depressive disorder. This is probably due to the Audio/Visual Emotion Challenge Workshop (AVEC) competitions, which provide automated extracted audio and video features to predict the severity of these conditions. Many other studies then used the public datasets in competitions like Distress Analysis Interview Corpus Wizard of Oz (DAIC-WOZ). Of the 264 studies in this review, 39% used DAIC-WOZ or AVEC datasets. Majority of the studies used some form of cross-validation for evaluating the performance of the trained models. However, only some studies used held-out test sets, which means that most models' reported performance may not generalize well. Without a held-out test set, performance may drop from the development set to the test set, as has been observed in AVEC competitions [32, 139, 140]. In contrast, models that used held-out test sets generally performed better on the test set [7].

Table 2.5: Exploring the Predictive Relationship Between Facial Expressions and Depression

Study	Main Findings
[127–129]	A deep residual regression model to evaluate depression levels using enhancement techniques can reduce the influence of external factors on the image, significantly improving prediction performance.
[130]	Proposed Part-and-Relation Attention Network for depression recognition, which outperforms state-of-the-art models with smaller prediction errors and higher stability.
[133]	Designed a model for depression detection using electroencephalogram (EEG) and facial features. A hybrid model is proposed, outperforming existing diagnosis systems.
[131, 135, 136]	A multimodal model with high performance on the AVEC 2013, AVEC 2014, and Emotion-Gait datasets. They concluded that the visual model is accurate.
[134]	An aggregation method which achieved comparable performance to 3D models with fewer parameters. The study suggests that video stimuli can be used for automatic depression detection.
[132, 137, 138]	Significant differences were observed in facial landmark features (e.g. average right nose (speed), median left ear top (speed), and left pupil-right pupil positions), and uniformed local binary pattern between healthy and depressive volunteers.

2.3.1 Optimizing Data Collection Strategies: Recommendations and Future Directions

The datasets used in automated depression detection studies vary greatly in size, participants' demographics, depression rating scales, the task used to elicit emotion, and the interview environment. Therefore, the performance of a detection model can be misleading if the dataset used for training is not representative of the studied population. In this review, we discussed the data collection strategies used in these studies to elicit emotions, record videos, and maintain participant privacy while avoiding confounding factors such as clinician questions and patient responses.

2.3.2 Identifying the Presence of Comorbidity

Many previous studies have not reported comorbidities [141] which presents additional challenges when developing automated depression detection models. However, Scherer et al. discovered a strong association (Pearson's $r > 0.8$) between scores for depression on the PHQ-9 scale and scores for PTSD on the PTSD Checklist – Civilian Version (PCL-C) in the DAIC-WOZ dataset [142]. Only a few articles stated that they excluded individuals with comorbidities from their study, such as Pan et al. [143]. To improve the data quality through consideration of comorbidities, researchers should use multiple mental disorder rating scales when collecting data. Additionally, researchers should develop and compare models trained with and without comorbidities to better understand the impact on the model performance.

2.3.3 Factors to Consider in Recruiting Control Groups

When selecting control groups for depression studies, it is important to ensure that individuals in the control group do not match any diagnostic criteria for other pathological conditions. For example, an individual may not be assessed as having depression based on their depression rating scale, but they may be assessed as suffering from PTSD that affects their speech patterns. Age, biological sex, first language, comorbidities, brain injury, respiratory disorders, and drug abuse can also affect speech and facial landmark patterns. In addition, variables such as education level, race, medication, and biological sex [144–146] can also affect speech patterns. Hert et al. have reported that antipsychotic therapies may lead to dyskinesia, an involuntary movement of facial muscles that affects speech and facial landmarks [147]. Therefore, individuals with a history of antidepressant medication should be excluded or reported in depression detection studies. Other variables such as biological sex, age, and education level can be adjusted via propensity score matching if they are statistically different between the depressive and healthy control groups.

2.3.4 Self-report Depression Rating Scales: Pros and Cons in Depression Diagnosis

The traditional method for diagnosing depression is via a clinical evaluation by a registered psychologist, which is considered the gold standard compared to self-report depression rating scales. However, clinical diagnosis can be costly and subject to the experience and expertise of the clinician, leading to lower inter-rater reliability [50]. Most studies included in this review relied on self-reporting depression rating scales instead of clinical evaluation, such as AVEC [145, 146] and DAIC-WOZ [144]. When using these self-rating scales, the task becomes predicting the self-report rating scale rather than a clinical diagnosis, which may not align with a clinician's evaluation. On the other hand, using open-source datasets for research can improve reproducibility and objectively compare model performance.

2.3.5 Eliciting Emotions in Depression Diagnosis

Choosing appropriate tasks for eliciting emotions is crucial, as specific features may be linked to certain depression rating scales but not others. In Table 2.6, we summarize the tasks used in previous articles and their advantages. Kane et al. proposed that sustained vowels are optimal for estimating glottal source features because it can be difficult to identify voiced sections in free speech [148]. Scherer et al. demonstrated that the voices of participants with moderate to severe depression are tenser than those of healthy participants [149]. Alghowinem et al. proposed that spontaneous speech leads to better results for most features than reading speech and that the first few seconds of speech perform better than the entire recording [150]. Another interesting approach to emotion elicitation is to use virtual agents for interviews, which can reduce data collection costs and can be less stressful for participants when discussing their symptoms. Multiple articles have reported successes in developing virtual interviewers [151–153], and the widely used AVEC challenges have also adopted virtual

interviewers.

2.3.6 Diarization of Speech Segments in Interview Recordings: Methods and Considerations

It is common practice to separate the speech segments of the participants from interview recordings to train depression prediction models. This process is commonly referred to as diarization. The participant's speech can be extracted using a microphone next to each speaker. If participants have headsets with lapel microphones during the interview, their voiced sections can be easily extracted, which may make some participants uncomfortable. Desk microphones can also be used, but they can introduce confounds because they are not targeted and make it difficult to separate participants' speech in the data processing. We suggest using two desk microphones next to each speaker with a sound barrier between them. It is important to record all metadata after the interview in a separate spreadsheet, such as participant ID, group, task, and other demographic information.

2.3.7 Ensuring Privacy in Interview Recordings

Clinicians should obtain verbal or written consent from the participants before interviewing. Participants must be informed that their interview recordings and demographic information may be distributed, pre-processed, and used for training machine learning models for academic research purposes. Even if participants grant permission for their data to be further processed, researchers must minimize the risk of data leakage because the interview audio (or video) recordings may contain sensitive information. To address this, researchers can share only the automated extracted speech and facial features rather than the raw interview recordings. If hackers were to gain access to the interview data, it would be impossible to reconstruct the original interview recordings using only the automated extracted features. Additionally, researchers can train the depression prediction model in real-time or use bone conduc-

tion microphones, which only record acoustic features without speech content [154], but this limits researchers to training semantic models. Edge computing can also be a solution to improve privacy by allowing computation to be performed on the participants' devices, with only the trained models being returned to the researchers, not the data.

2.3.8 Data Preprocessing and Automated Feature Extraction

Automatic speech recognition (ASR) can transcribe speech into transcripts for training semantic-based depression prediction models. ASR can also filter unvoiced sections and noise in the interview audio (or video) recordings. In most in-person interviews, two speakers are present, and the clinicians' segments can be discarded if the ASR system includes automatic diarization. To prevent overfitting, techniques like dimensionality reduction or feature selection should be applied to the training and test sets during preprocessing. This will help ensure that the model is not overly influenced by the specific characteristics of the training data and can be generalized to new data.

The most commonly used automated feature extraction tools in the studies we reviewed were openSMILE, COVAREP, pyAudioAnalysis, and openEAR. To ensure the deep learning models converge, it is recommended to standardize or normalize features as they may be in different scales. Before training the model, we recommend performing exploratory data analysis or visualization to better understand how these features characterize mental disorders. This can help inform the selection and preprocessing of features, as well as the design of the model.

2.3.9 Evaluate Models with Small Datasets: Bootstrapping and K-fold Cross-validation

To avoid overfitting the model on the test set, we typically evaluate the trained model on the held-out test set only once. However, when training a model for predicting de-

Table 2.6: A Comparative Study of Different Speech-eliciting Tasks

	Task and examples	**Advantages**
Constrained	Repeating "PATAKA" [64, 155]	Capture speech sequencing; A proxy for lung capacity
	Sustained vowel [156]	Measure muscle weakness and aspects of move control
	Counting [157]	Counting from 1 to 10 allows mroe control over acoustic patterns
	Reading	
	• The "Nordwind" passage [30, 158]	Paragraph frequently used in the gathering of depression-related speech
	• Rainbow passage [156]	Includes all the sounds used in English and reflects normal speech patterns
	• Emotion-evoking movie clips [159]	Greater ability to regulate emotions that are provoked
Free speech	Monologue	
	• Describing, memory recalling [160]	More spontaneous than reading speech
	Dialogue	
	• Semi-structured interviews [144]	Frequently used in medical facilities
	• Phone conversations [161]	Only the interviewee is recorded; no need to identify the speaker

pression using a small dataset (e.g. around 100 data points), which is commonly seen in the medical field, a 20% held-out test set or K-fold cross-validation can decrease the number of samples available for training. In addition, a small test set is unlikely to represent the entire population accurately. As a result, we suggest using repeated bootstrapping to evaluate the depression prediction model, which provides a distribution of performance metrics with mean and standard deviation. However, given the computational complexity of deep learning models, the bootstrapping method may not be feasible. When working with a small dataset, K-fold cross-validation can be a viable alternative to bootstrapping, as deep learning models tend to need a large number of data points which reduces the need for bootstrapping.

2.3.10 Evaluating the Performance of Depression Prediction: Best Practices and Considerations

Performing better than chance does not indicate the model learned from the training data, and the resulting metrics must be generalizable and statistically significant for clinical use. Aloshban et al. demonstrated that their accuracy is always better than chance to a statistically significant extent [162]. However, to further prove generalizability, we suggest performing a permutation test in future works, where models are trained on permutated labels to evaluate the model's performance based on mistaken labels, which is often better than chance. A statistical test can then determine if the difference between the permuted and non-permuted scores is statistically significant. On the other hand, clinical datasets can be imbalanced, with a greater number of healthy cases compared to the population of individuals with depression. In this case, using the accuracy of the classification model as the sole metric to evaluate its performance may not be objective since it will be biased towards predicting every sample as negative. To evaluate model performance more objectively, metrics such as the F_1 score, precision, recall, and area under the curve (AUC) should be considered. These metrics account for class imbalance and provide a more balanced view of model

performance. Saito et al. have shown that the precision-recall curve is more useful than the receiver operating characteristic curve when evaluating binary classifiers on imbalanced datasets [163]. In addition to the precision-recall curve, metrics such as root mean square error (RMSE), mean square error (MSE), and the coefficient of determination (r^2) are commonly used to evaluate the performance of regression models for predicting depression scores. In the AVEC competition, the performance of baseline models was evaluated using the concordance correlation coefficient (CCC), which takes into account changes in scale and includes measures of both precision and accuracy [164]. It is generally helpful for other researchers to see a range of metrics when evaluating the performance of a model, as this allows for more objective comparison. A model's performance must be generalizable and statistically significant to be truly useful in a clinical setting.

2.3.11 Explainable Depression Detection Model

Recent evidence suggests that individuals may lack trust in black box models and that these models may cause harm in high-stakes decision-making processes [165–167]. As a result, researchers are exploring ways to explain the decision-making processes of algorithms better [168–170] through publications and software packages implement explainable machine learning models [170, 171]. By providing explanations of a model's feature contributions, clinicians can gain a better understanding of depression and improve the model itself. Once high-impact features have been identified, we can retrain the model using only these features to evaluate their performance. In some of the reviewed articles, we observed that while the studies presented excellent feature engineering for distinguishing between groups, they lacked quantitative analysis to support their findings. We suggest linking changes in the automated extraction of features to mental disorder symptoms to provide a more comprehensive view of the model's performance.

2.3.12 Ensuring Reproducibility in Automated Depression Detection

Reproducibility is a critical issue in machine learning, particularly when artificial intelligence is applied to healthcare [172, 173]. One obstacle to reproducing previous studies is that clinical datasets are not always available for redistribution. As we mentioned in Section 2.3.7, automated extracted features can be shared without violating privacy concerns. However, sharing the code used for training and evaluating the model is also important. Even when the code and data are publicly available, other researchers may still have difficulty reproducing the results due to differences in the software environment. To address these issues, we suggest that researchers use containers, such as Docker, which include the data, code, and environment in one package that can be easily redistributed. This will make it easier for other researchers to reproduce the results, ultimately accelerating the advancement of automated depression detection.

2.4 Future Work

2.4.1 Ethical Considerations in Automated Depression Detection

Automated depression detection can benefit society by reducing the workload of the healthcare system, preventing suicidal or self-harm behaviours, and enabling law enforcement authorities to track abnormal behaviours. However, the use of automated depression detection also raises some ethical concerns. For example, insurance companies and employers may use the results to evaluate candidates without their knowledge or consent and reject them if a mental disorder is present or likely to develop in the future. Additionally, it can be difficult for individuals to fully understand the implications of consent forms, which can further complicate the ethical considerations surrounding automated depression detection [174]. To ensure that automated depres-

sion detection is used ethically in clinical settings, researchers should provide clear and understandable explanations of how the collected data will be used. Participants should also have the right to revoke permission to use their data at any time. Like other developing technologies, these systems may be vulnerable to abuse and have unexpected side effects. As researchers, engineers, and clinicians, it is our responsibility to educate the public and policymakers about the potential benefits and harms of automated depression detection to both prevent abuse and further advance these techniques, which have the potential to help many people.

2.4.2 Leveraging Machine Learning for Advancing Psychiatry

With this chapter, we aim to demonstrate the potential for psychiatry to benefit from advances in machine learning. Many individuals have difficulty accessing qualified mental healthcare or may be hesitant to seek psychotherapy due to stigmatization [175]. Automated depression detection models can provide an accessible and efficient method for early screening, which can help individuals determining that they may need professional healthcare. Additionally, psychiatric visits often include interviews that can be recorded in video or audio format, which provides a wealth of data that can be used to associate mental health assessments with acoustic, semantic, and facial features. By following the guidelines outlined in this section for collecting and analyzing this data, we hope to enable new collaborations between clinicians and machine learning engineers to advance our understanding of mental health disorders.

2.5 Conclusion

We reviewed 264 studies that measure acoustic, semantic and facial landmark features to distinguish between individuals with and without mental health disorders using either null hypothesis testing or predictive machine learning models. Our synthesis in-

cludes significant and non-significant features across audio, text and facial modalities, as well as those correlated with the severity of depression. We also provide guidelines on collecting data, preventing confounding factors, protecting privacy, selecting speech-eliciting tasks, and improving machine learning model generalizability and reproducibility. We also found a few studies have been conducted on post-traumatic stress disorder, bipolar disorder and postpartum depression, thanks to open-access research datasets provided by the AVEC and DAIC. Based on their proven effectiveness, we encourage the collection of open datasets, particularly distributing datasets through competitions. These are highly productive in advancing research in various fields. While productivity is important, reproducibility is also critical. Since the studies in this review involve building computational models, the associated data and code should be shared, ideally through containers. This allows others to test the claims made by these studies and contribute to the development of these models in a collaborative manner. Moreover, conducting more research on multiple datasets may help enhance the models' generalizability and reconcile conflicting results regarding crucial and predictive features. This approach could lead to more robust and reliable conclusions about the nature of these disorders and their diagnosis and treatment. Using multimodality features to train machine learning models holds promise for enhancing mental health evaluations and treatment. This approach aligns with the principles of preventive and personalized diagnosis and treatment and could lead to better outcomes for individuals with mental health conditions.

Chapter 3

Prediction of Depression Severity Based on the Prosodic and Semantic Features with Bidirectional LSTM and Time Distributed CNN

We aim to address some limitations and gaps identified in previous studies. Based on the insights gained from the literature review, we introduce a novel multimodality automated depression diagnosis system that leverages prosodic and semantic features. One advantage of our proposed model is its flexibility regarding input duration. Unlike previous approaches that required input sequences to be of the same length, our model can handle audio/text feature sequences of varying lengths. By bridging the gap between existing research and our proposed methodology, we aim to contribute to the advancement of automated depression diagnosis systems.

3.1 Introduction

Mental health disorder, such as depression, is considered one of the major challenges facing global society. During the COVID-19 pandemic, the prevalence of depression and anxiety is exacerbated in the general population [176–179]. By 2030, depression will be the second major cause of disability worldwide and thus it can impose a heavy

healthcare burden globally [180]. However, often the symptoms of depression are not displayed directly. Many individuals often express their sadness and hopelessness but without depression, whereas patients are usually reluctant to report their conditions and receive treatment [181]. For instance, many people with depression ignore or refuse to admit their emotional instability and physical health conditions. The reason is that depression is a stigmatized disease, resulting in the depressive population hiding or camouflaging their symptoms. Traditionally, a semi-structured clinical interview based on Diagnostic Statistical Manual (DSM) criteria is the standard protocol for depression diagnosis [182] with self-test questionnaires such as the Patient Health Questionnaire Depression Scale (PHQ) [183], Beck's Depression Inventory (BDI) [184] and Montgomery-Asberg Depression Rating Scale (MADRS) [185]. The PHQ-8 is an assessment form created to examine the existence of core depression symptoms, such as fatigue and anxiety. The PHQ-8 scale shows high sensitivity and specificity for diagnosing depression and other mental disorders among patients with different languages and cultures[186]. These methods play a key role in diagnosing depression, but the results are subject to physicians' experience. Previous articles argued that these clinical criteria, such as DSM and BDI, are not reliable enough [187]. Diagnosis of depression is not the same as other medical conditions since gold standards for mental disorders do not exist currently, which raises the likelihood of misdiagnosis and finally leads to unexpected results [115, 188, 189]. However, most depressed people do not have access to qualified psychological treatment due to economic conditions (low-/mid-income population) or living constraints (in rural regions) [190]. Therefore, it will be beneficial to develop a low-cost screening technique that can be deployed in communities and operated by people without special training. Early-stage mental disorder screening is also crucial for policymakers and security agencies because someone with a mental health disorder could behave adversely to other innocent people, such as massive shootings which are attributed to mental

health disorders [52]. In the cyber world, we live now, it is very common to share personal information, and concerns through the Internet, especially after the rise of social media. This raises an opportunity since the contents on social media increase the likelihood of detecting potential depression patients from a large population.

In this chapter, we propose a multimodality automated depression diagnosis system with prosodic and semantic features to predict depression levels with the combination of Bi-LSTM and TD-CNN models. To the best of our knowledge, it is the first time that time-distributed CNN is adopted to further extract the temporal information from the output of the LSTM encoder. Additionally, our proposed model does not have a strict limitation of input duration, regardless of the number of frames, as long as the number of features meets our specification, our model can always provide a patient-independent depression prediction. The prediction is based on a specific text or audio feature sequence. Given a specific participant with an audio/text feature sequence of arbitrary length, our model provides a series of estimations of depression severity based on the audio/text feature. The set of predictions can be merged through a major voting algorithm so that the final output of our model is a patient-level depression severity prediction. This mitigates the problem that the audio/text feature sequences are required to be the same in length in previous articles. LSTM performs well in learning temporal information because of its recurrent structure. The bidirectional LSTM model is used to learn long-term bidirectional dependencies in the audio and text feature sequences because it has been proven to perform better than a unidirectional LSTM model. The convolutional neural network (CNN) is a popular network architecture for learning the spatial features of data. A time-distributed CNN architecture is obtained by having multiple CNN layers for Bi-LSTM output features at each timestep. Given the complementary advantage of CNN and LSTM, the hybrid model of LSTM and TD-CNN works well in learning the spatiotemporal sequence. The best patient-independent F_1 score of the audio and text model is 0.9870

and 0.9709, respectively, on the test partition of the DAIC-WOZ dataset. The fused multimodality model achieved the best F_1 score of 0.9580 on the test partition of the DAIC-WOZ dataset. This chapter has been published as "Prediction of Depression Severity Based on the Prosodic and Semantic Features With Bidirectional LSTM and Time Distributed CNN," in IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 2251-2265, 1 July-Sept. 2023, doi: 10.1109/TAFFC.2022.3154332.

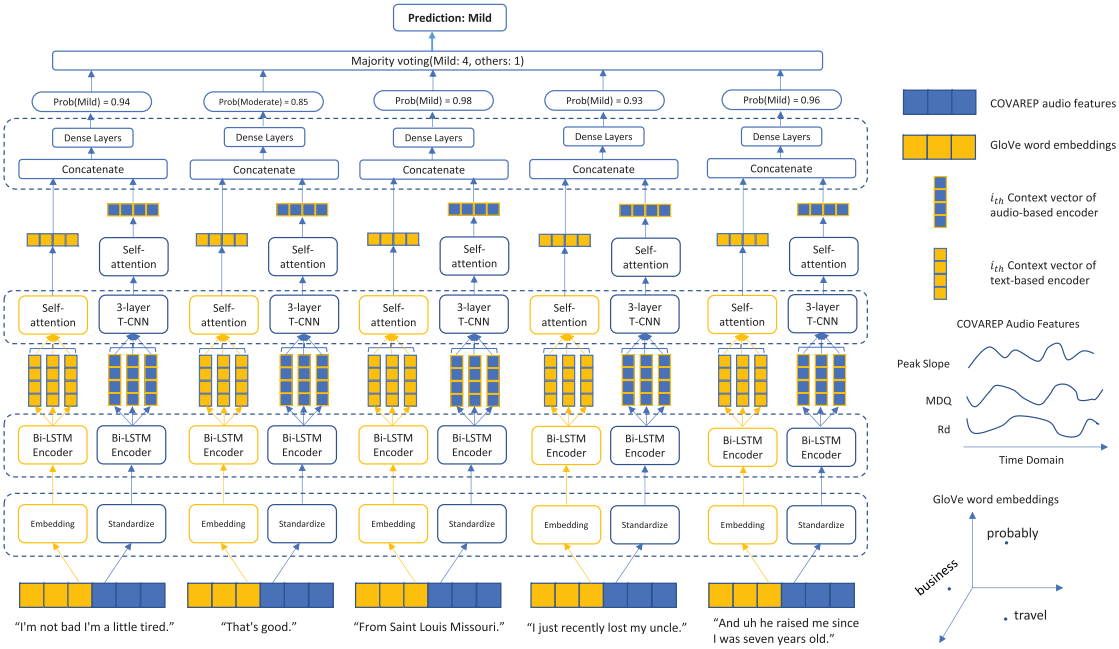

Figure 3.1: Block diagram of our proposed multimodality depression level prediction algorithm given a specific example. Audio features are fed into the network through the input layer. After batch normalization, the input data is fed into the Bi-LSTM and time-distributed CNN block. In this proposed design, we have five time-distributed CNN blocks followed by a single-layer Bi-LSTM. The detailed architecture of each block is illustrated and explained in the remainder of this chapter.

3.2 Methods and Procedure

In this section, we briefly introduce the preliminary material we used for developing the audio model, text model, and multimodality model. We also discuss the dataset and framework for training and evaluating our proposed model.

3.2.1 Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ)

We adopted the Distress Analysis Interview Corpus-Wizard-of-oz (DAIC-WOZ) dataset for training and testing. The corpus consists of 189 recorded clinical interviews and transcripts as well as facial features from 189 subjects. The audio recordings were taken of semi-structured interviews between the participants and a virtual interviewer called Ellie, an animated role controlled by a human interviewer. The average audio duration of 189 subjects is 974 seconds. Subjects were solicited from the Greater Los Angeles Metropolitan region from two different populations. One was from civilians; the other was from veterans of the U.S armed forces. Subjects were characterized as depression, Post-Traumatic Stress Disorder (PTSD), and anxiety based on the self-report questionnaire during the data collection [144]. Only the interview recordings of the depression group were released for academic purposes. The gender (biological sex) distribution over all five groups as well as the dataset partition is shown in Table 3.2. In the training set, there are 44 female subjects (27 without significant depression symptoms, 17 with depression symptoms) and 63 male subjects (49 without significant depression symptoms, 14 with depression symptoms). In the validation set, there are 19 female subjects (12 without significant depression symptoms, 7 with depression symptoms) and 16 male subjects (11 without significant depression symptoms, 5 with depression symptoms). In the test set, there are 24 female subjects (17 without significant depression symptoms, 7 with depression symptoms) and 23 male subjects (16 without significant depression symptoms, 7 with depression symptoms). All interviews were transcribed verbatim into English. The interviews lasted from 5 to 20 minutes involving three phases: it started with neutral questions, which aimed to ensure subjects being able to calm down; the interview then proceeded into a targeted phase, and the questions asked by the interviewer were more related to the symptoms of depression and PTSD. Finally, the interview terminated with the

annealing phase, which assisted the participants to get rid of the distressed state. The PHQ-8, ranging from 0 to 24, determines the severity of the mental disorder. Subjects were divided into five groups: healthy (PHQ-8<5), mild (5<PHQ-8<10), moderate (10<PHQ-8<15), moderately severe (15<PHQ-8 <20), and severe (PHQ-8 >20) [191]. Table 3.1 shows a sample transcript in the DAIC-WOZ dataset, which contains four fields: beginning and end timestamp of the utterance, the speaker ID, and sentence content. In the remaining part of this chapter, the training, validation and test set are split by the instruction from the DAIC-WOZ dataset independently, which ensures all the subjects only appear in one of the above partitions.

Table 3.1: The Showcase of a Participant's Transcript

Start time	Stop time	Speaker	Utterance
87.322	89.592	Ellie	So how are you doing today?
89.71	91.93	Participant	I'm not bad I'm a little tired but okay.
92.945	93.585	Ellie	That's good.
94.257	95.577	Ellie	Where are you from originally?
95.78	97.14	Participant	Uh from Saint Louis, Missouri.

3.2.2 Audio Features and Models

In this section, the audio features are extracted by COVAREP [192], which can be divided into three categories: glottal flow features (NAQ, QOQ, H1-H2, PSP, MDQ, Peak slope, Rd), voice quality features (F_0, VUV), and spectral features (MCEP,

Table 3.2: Gender (biological sex) Distribution Over All Groups and Dataset Partitions

Dataset profile for depression level classification						
	Female	Male	Female	Male	Female	Male
#Healthy	7	9	2	3	3	2
#Mild	12	25	7	6	9	11
#Moderate	10	20	5	2	7	3
#Moderately severe	10	5	2	2	1	4
#Severe	5	4	3	3	4	3
Dataset profile for depression detection						
#Subjects w/o significant symptom (PHQ-8≤10)	27	49	12	11	17	16
#Subjects w/ significant symptom (PHQ-8>10)	17	14	7	5	7	7

HMPDM, HMPDD). Normalized Amplitude Quotient (NAQ) quantifies the time-based feature of the speaker by amplitude-domain measurements calculated from the glottal flow and its first derivative [193, 194], Quasi Open Quotient (QOQ), which is a correlate of the open quotient (OQ) which involves the derivation of the quasi-open phase based on the amplitude of the glottal phase [195, 196], the amplitude difference of the first two harmonics of the differentiated glottal source spectrum (H1H2) [193], Parabolic Spectral Parameter (PSP), which is based on the quantification of the spectral decay of the speaker [197], and Maxima Dispersion Quotient (MDQ), which is designed to quantify the maxima dispersion as a result of phonation type moves towards a breathier phonation [198, 199]. Spectral features consist of Mel-Cepstral Coefficients (MCEP0-24), which is a representation of the short-term power spectrum of a sound [32], harmonic model and phase distortion mean (HMPDM0-24) and deviation (HMPDD0-12). Thus, there are 74 audio features in total. Each subject is represented in the COVAREP features, $X_i \in R^{T \times F}$ where T denotes the time dimension, which is proportional to the duration of the audio. Each 10 milliseconds frame of audio was transformed into an audio feature vector. F denotes the number of features

COVAREP extracted for each frame. Among the 74 audio features, the entry “VUV” indicates whether the audio features are extracted from the audible or silent part of the original interview recording. Only those audio features where ”VUV” is 1 can be the input to the following models. Among all the 189 subjects in the dataset, audio features are in an average of 35850 frames (rows) and a standard deviation of 15791 frames (rows). For each subject, we concatenated a constant number of audio feature frames into a set of successively retrieved audio feature sequences, which were used to represent this subject. The shape of the input tensor is thus (#samples, #frames, 73). The field ”VUV” is always 1 in the input tensor so it is dropped, which results in the final input tensor shape as 73.

Audio models with different configurations for depression assessments are introduced as follows. The input to these models is the previously mentioned audio feature sequences, the output of these models is the prediction of the depression severity given an audio feature sequence. The first audio model is a simple one that consists of the LSTM and fully connected layers. The LSTM served as a feature extractor and the following fully connected layers made the prediction based on the output of the LSTM. Then, we introduce our proposed model that consisted of the Bi-LSTM and TD-CNN and they were evaluated for the prediction of depression severity.

Traditional LSTM-based Model

Our first audio model comprises of single-layer Long-Short Term Memory (LSTM) network and fully connected layers. LSTM network was obtained using an LSTM layer containing 73 hidden units, connected to a fully connected layer. To avoid overfitting, the dropout was applied to the recurrent input signal on the LSTM units and between fully-connected layers with the dropout rate of 0.2. The time step is equal to the constant ”#frames” and there were 73 features in each timestep. In this model, only the hidden state at the last time step was fed into the following fully connected layers, with 128 and 64 hidden units. The output of the fully connected

layer was then fed into a batch normalization layer and flattened into a 1D tensor. The flattened tensor was fed into a fully connected layer with 5 hidden units, where the SoftMax activation function transformed the unnormalized output of each neuron into the probabilities of five severities. An Adam optimizer was adopted for the training, the initial learning rate was set to be 0.001, β_1=0.9, β_2=0.999 and the epsilon was 10^{-7}. A callback function monitored the validation loss and terminated the training if the validation loss did not decrease after five epochs. A loss function of cross-entropy was applied.

Hybrid of Bi-LSTM and TD-CNN Model

Bidirectional LSTM is a variant of LSTM which consists of a forward layer on the original input sequence, and a backward layer on the reversed sequence. The Bi-LSTM outperforms the traditional LSTM because the forward and backward networks combine both forward and backward context information of the input sequence. Previous articles proposed to represent the input sequence by the last hidden state of the LSTM [32, 200]. However, depression assessment is a complicated task, which heavily relies on the relationship between the audio features at different time steps, thus it is insufficient to use the last hidden state for classification, otherwise, it leads to the loss of temporal information. To solve this issue, we utilized the TD-CNN (shown in Figure 3.2) to learn potential temporal and spatial information in the output of the Bi-LSTM. In general, simple CNN only supports the 2D or 3D spatial tensors as the input. However, the output shape of the LSTM is (#samples, #frames, #LSTM neurons) given a unidirectional LSTM, and (#samples, #frames, 2*#LSTM neurons) given a bidirectional LSTM. The TD-CNN convolves the LSTM output vector along its 3rd axis and the shape of the convolution result is (#samples, #frames, #output features, #kernels). Therefore, we expand the shape of the LSTM output vector by inserting one new axis so that it can be processed by TD-CNN. The TD-CNN accepts a tensor with shape (#samples, #frames, 2*#LSTM neurons, 1) as the input, which

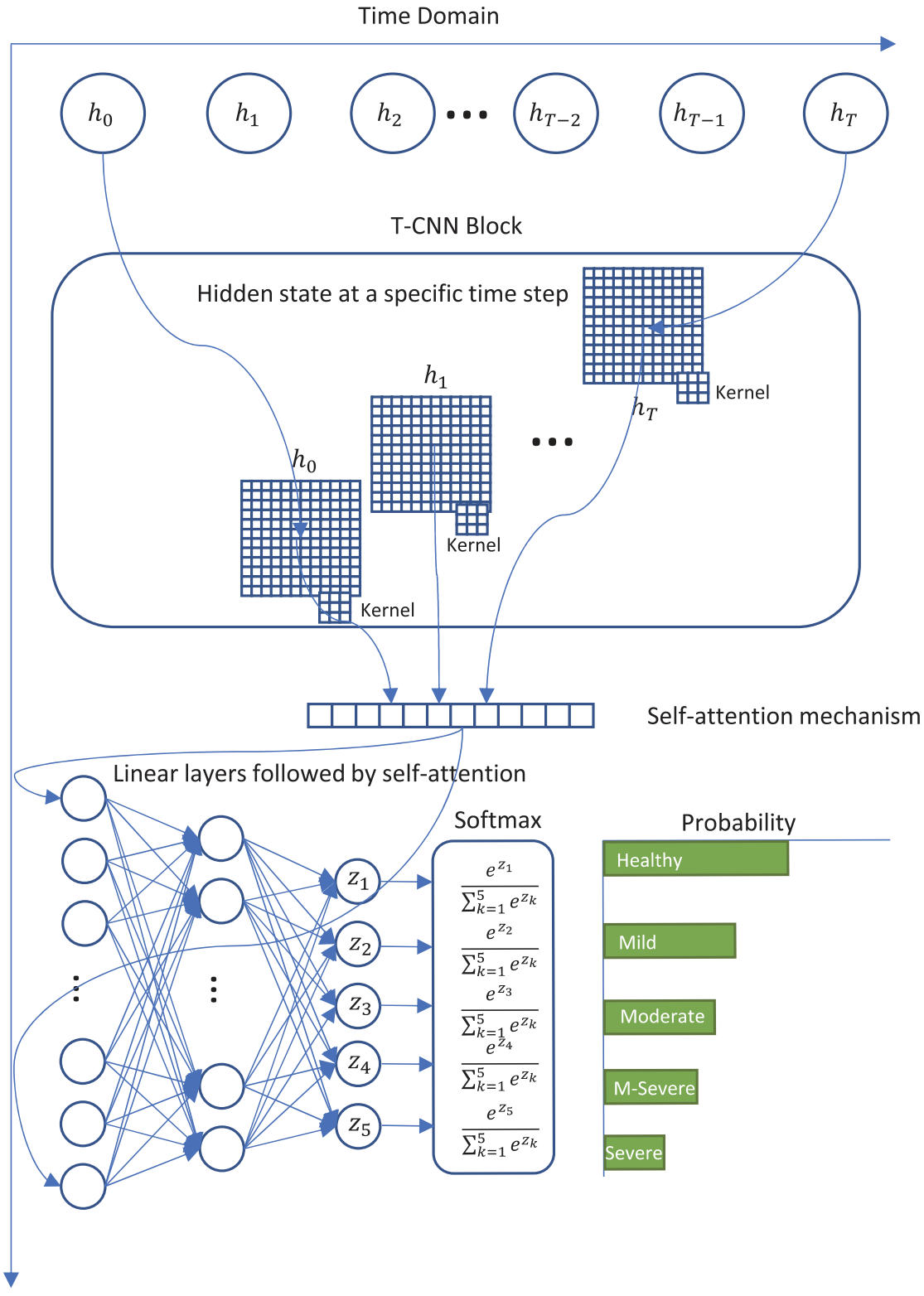

Figure 3.2: The structure of the TD-CNN model and the following linear neural network.

denotes a time series of LSTM hidden states. Our proposed TD-CNN block consisted of three layers, first, time-distributed convolution layer, then time-distributed pooling layer to downsample the feature maps; and finally batch normalization layer. There were five TD-CNN blocks in total in our proposed design, the output of the last TD-CNN block contained "#frame" samples, each sample is represented by 256 feature maps. Therefore, the last thing before the feature maps were fed into the following network was to downsample the output by the global average pooling layer, it slides along the time dimension of the feature vector and computes the mean value of each feature, which ensures that the relationship between each time step was taken into consideration. The output of the global average pooling layer was then fed into the following two linear layers. At last, the Softmax activation functions transformed neuron output into the probability of five severities. An Adam optimizer with a similar configuration in Section 3.2.2 was adopted for the training.

3.2.3 Text Features and Models

The input layer of the text model took tokenized transcripts of each subject. Among all the 189 subjects in the dataset, text transcripts are in an average of 80 rows and a standard deviation of 14 rows. The interviews were in colloquial speech, thus the first step was to rephrase these colloquial descriptions to written languages, otherwise, colloquial terms all became out-of-vocabulary words, which were represented by the token [UNK], and greatly diminished model performance.

Semantic information is highly essential in depression diagnosis because psychologists also formulate diagnosis by text produced by the patients during the interview. To acquire the text features, we firstly removed stop words in the patients' responses with Natural Language Toolkit (NLTK) and substituted some words and phrases such as "what's", "e-mail" with "what is" and "email", this eliminates different expressions of the same word [201].

Next, we lemmatized the remaining words in the sentences, the WordNet lemmatizer removes the inflectional endings and returns the base form of a word. Then the remaining texts were tokenized into word lists and were used to build a vocabulary with 7373 words. Each word in the vocabulary was assigned an index, the word list was then represented by these indices. After we acquired the word list, the main issue was that each word list was different in length, which made it more difficult to batch process text data if they were different in length. Therefore, the sliding window technique was applied to generate sequences in the same length, which was the same length as the sliding window. Each window consists of a constant number of words while 20% words at the end were overlapping between two neighbouring time windows, which assigned higher weights to the words at the edge of the window so that the edge details were enhanced. The sliding window not only generated all training pairs but also performed data augmentation as well as directed the focus on a specific part of the sentence. Next, word sequences were encoded with the pre-trained 100D GloVe word embedding vector [202]. The word embeddings were concatenated into a sentence embedding. For some short sentences, the size of the sliding window was greater than the length of the sentence, those short sentences were zero-padded to be the same length as the window. Therefore, the shape of the final input vector is (window size, 100). However, sentences shorter than 20% of window size were discarded.

Bi-LSTM Text Model

Our proposed text model consists of a single-layer Bi-LSTM network and fully connected layers. The text feature sequences mentioned above comprise the index of words in the vocabulary. Text feature sequences were preprocessed to map each word to word embedding space with a non-trainable embedding layer before being fed into the model, and the shape of the embedding layer is (vocabulary size + 1, 100). Next, a batch normalization layer and then the Bi-LSTM layer further captured the se-

mantic information underlying the input word sequences. To avoid overfitting, the dropout was applied to the recurrent input signal on the LSTM units and between fully-connected layers with the dropout rate of 0.2, and the shape of the Bi-LSTM output was (batch size, 200) at each time step. We adopted the attention mechanism to allow the model to adaptively select those depression-sensitive hidden states. The attention vector was then fed into two linear layers with 256 and 128 hidden units, respectively. Finally, the last linear layer with 5 hidden units determined the probability of the five severities. An Adam optimizer with a similar configuration in 3.2.2 was adopted for the training. The cross-entropy loss calculated the distance between the output and the ground-truth label.

3.2.4 Fused Text-Audio Joint Model

Our final fused multimodality model was comprised of two sub-networks: text model and audio model, and followed by a shared late fusion neural network as Figure 3.1 shows. The late fusion neural network concatenated the outputs of the text and audio model to integrate text and audio features. For any subject, we extracted a high-level representation that included both semantic and prosodic features through the previous recurrent neural network and convolutional neural network. This high-level representation could be used in the following assessment of mental disorders. The output of our proposed model was a scoring matrix that denoted the likelihood of the depression severity. As the timesteps of the audio and text model were different, the late fusion network had to deal with input of different sizes. To solve this issue, we first attempted to adopt a max-pooling method to downsample the output from audio and text models so that they were in the same shape. Moreover, an attention mechanism was exploited, which provided us insights into the ratio of the contribution of each modality towards the final prediction.

Regarding fusion, we designed a set of models to integrate different modalities.

Firstly, we fused the text models with different window sizes with the audio model with constant configuration. Our text model could be divided into two categories, one is the unidirectional LSTM text model, the other is the bidirectional LSTM. Our proposed audio and text model was previously described in Section 3.2.2 and Section 3.2.3, respectively. The only difference was that the output size of the audio and text model was 32 instead of 5 since they acted as feature extractors rather than classifiers. Global max pooling was adopted to align the extracted audio and text features. In order to integrate text and audio modalities, the output of the text and audio model was concatenated into a tensor and passed through a fully connected layer with 5 units. Secondly, the other fused model was set up using a similar configuration to the first one. The difference was that the attention mechanism played its role in aligning the features from different modalities. The third one was all the same as the previous two models, except it was created with an attention mechanism not only during the feature alignment but also in the fusion of the high-level representations.

3.3 Results and Discussion

In this section, the results of those models described in Section 3.2 are presented and discussed. We next assessed the effect of the hyperparameters for the proposed models. For the audio model, we compared the effect of architecture and timestep and investigated the potential long-term dependency of the audio features in severe patients. For the text model, we conducted experiments to investigate the effect of the hyperparameters such as the size of the window in preprocessing, the removal of stop words. Regarding the audio-text fused model, we mainly focused on the impact of fusion methods on the model performance. All the experiments were conducted on one RTX 2080Ti 11GB GPU. The size of multimodality models was limited mainly by the amount of memory available on our GPU and the amount of time for training we can tolerate. Our single-modality model usually took between 3 to 5 hours to

train, but the training of our proposed multimodality model always took around 20 hours. The results of our experiment provided an insight that our models could be improved by faster GPUs and larger datasets. The detailed results are discussed in the following parts.

3.3.1 The Statistics of Audio and Text Features

The pause time between responses is also longer than usual in the depressive population [55]. To verify whether the DAIC-WOZ dataset follows a similar pattern, we calculated the statistics of the raw interview recordings and the transcripts. The subjects were divided into two groups by PHQ-8 scale, the subjects were considered as normal or mild (control group) if their PHQ-8 is less or equal to 10, otherwise, they are considered as moderate or severe (experiment group). This threshold is given by a previous study on the efficacy of PHQ-8 on the diagnosis of major depressive disorder. It was reported that given the cutoff score of 10, the PHQ-8 exhibited a sensitivity of 58.3%, specificity of 83.1% [203]. The two-sided T-test was applied to test if there was a significant difference in the audio duration between the control and experiment groups. Additionally, Cohen's d was calculated to quantify the effect size of the observed differences between the groups. Cohen's d is a standard score that measures the the difference between the mean value of groups. Cohen's d is given by Equation 3.1. The statistics of the two groups are listed in Table 3.3. The histograms of the audio duration and sentence length of the control and experiment groups are illustrated in Figure 3.3. The response duration of the control and experiment groups is on an average of 951.37 $\pm$ 266.60 and 997.87 $\pm$ 290.19 seconds, respectively. The two-tailed p-value is 0.09 and Cohen's d is 0.17. The sentence length of the control and experiment groups is on average 8.78 $\pm$ 8.94 and 7.37 $\pm$ 7.29 in the number of words, respectively. The two-sided T-test was applied to test if there was a significant difference between the sentence length in the control and experiment groups. The

two-tailed p-value is 3.23×10^{-14} with Cohen's d at 0.16. The above results indicate no significant difference in the audio duration of the control and experiment groups. However, there is a slight difference in sentence lengths, with Cohen's d at 0.16. More responses in the experiment group consisted of less than five words. As the audio durations between the control and experiment groups have identical average values, we can conclude that there are more pauses in the conversations of the experiment group. This result is identical to other researchers' conclusions. Therefore, our dataset and criterion for depression are reasonable.

$$
\begin{aligned}
d &= \frac{\bar{x}_1 - \bar{x}_2}{s} \\
s &= \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}}
\end{aligned}
\quad \text{where} \quad
\begin{aligned}
&d : \text{Cohen's d (effect size)} \\
&s : \text{pooled standard deviation} \\
&\bar{x}_1, \bar{x}_2 \text{ are the means of two groups} \\
&n_1, n_2 \text{ are the sizes of two groups} \\
&s_1, s_2 \text{ are the variances of two groups}
\end{aligned}
\tag{3.1}
$$

Table 3.3: T-Test Result of the Control and Experiment Group

	Audio duration		Sentence length	
	Control	Experiment	Control	Experiment
Mean	951.37±266.60	997.87±290.19	8.78±8.94	7.37±7.29
p-value	0.09		3.23×10^{-14}	
Cohen's d	0.17		0.16	

3.3.2 Results of the Audio Modality

As for the audio models, evaluation metrics accuracy, recall, precision, and F_1 score used to evaluate models with different configurations are shown in Tables 3.4 and 3.5. The test set for evaluation is balanced by oversampling the minority class. Ran-

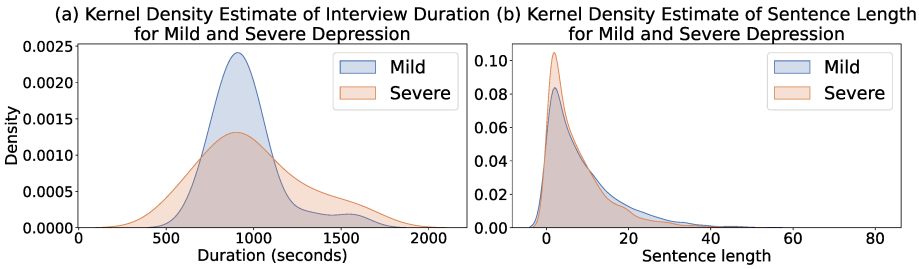

Figure 3.3: Kernel density estimations of the audio duration and sentence length of control and experiment groups. (left) The audio duration of the control and experiment groups. (right) The sentence length of the control and experiment groups.

Table 3.4: Results of the Baseline Audio Models

	Random Forest		[204]	[205]
	Mean	St. dev	Mean	Mean
Accuracy	0.3192	0.0085	0.7500	0.8273
Precision	0.3206	0.0064	0.7200	0.7930
Recall	0.3184	0.0040	0.7500	1.0000
F1 Score	0.3168	0.0076	0.7300	0.8850

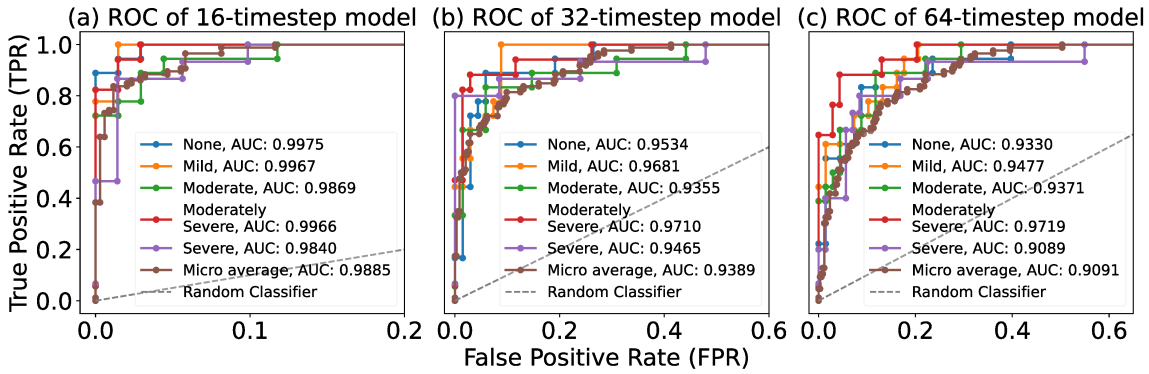

Figure 3.4: The ROC of three different model configurations. (left) The Bi-LSTM followed by TD-CNN given the time step = 16. Micro-Average AUC: 0.99. The AUC of "Severe" is smaller than any other class, this indicates the detection of severe depression is more challenging than other depression levels. (middle) The Bi-LSTM followed by TD-CNN given the timestep = 32. Micro-Average AUC is 0.94. The micro-average AUC is smaller compared with that when the timestep = 16. The longer sequence does not mean a better result because the noise introduced by the longer sequence can mislead the model. (right) The Bi-LSTM followed by TD-CNN given the timestep = 64. Micro-Average AUC is 0.91, which is in line with our expectation that a longer input sequence makes it more challenging to predict the severity.

dom forest was used as the baseline in evaluating the audio modality sequence-level prediction. Audio feature sequences for training and evaluating are non-stationarity series, which are difficult to model and forecast. They were pre-processed by differencing to be made stationary. Differencing is the change from one audio feature sampling time to the next. The random forest model we used in this manuscript is an ensemble approach that fits a set of decision trees on different sub-sample of the dataset, and averaging the output of each decision tree to improve the prediction accuracy, as well as prevent the model from overfitting. In our article, 100 decision trees were trained on various sub-sample of the training set to construct the random forest model. Another baseline method, Madhavi et al. proposed a CNN consisting of 2 convolutional layers and two successive linear layers to extract high-level features from the frequency spectrogram of interview recordings. The output of CNN is fed into the following neural networks to predict an individual's depression level. They also

evaluated their models on the DAIC-WOZ dataset. Moreover, Yang et al. proposed a similar but more complex model, they also adopted the combination of convolution neural networks and deep neural networks (i.e. multi-layer perceptron model). Each subject was labelled by their depression-related symptoms, such as prior depression diagnosis, sleep disorder, present or not. Their proposed CNN consists of three convolution layers and the intermediate output of CNN is fed into the deep neural network to predict the presence of depression symptoms. These symptom labels are fed into another deep neural network for predicting depression severity. Their results on the DAIC-WOZ dataset are summarized in our comparative studies. For the LSTM with the fully connected layers model, it outperformed the baseline machine learning model (i.e. decision tree) by 24% in terms of accuracy. In contrast, the Bi-LSTM with the fully connected layers model outperformed by 54% in terms of accuracy. For our proposed Bi-LSTM combined with the TD-CNN model, we achieved 16% improvements over the best baseline model in terms of accuracy. From Tables 3.4 and 3.5, it can be concluded that the LSTM performed better on the depression level classification compared with the baseline machine learning models, such as the naïve Bayes model. Moreover, we observed that the network followed by the LSTM layer is critical for good performance. If the other configurations were fixed, Bi-LSTM with TD-CNN outperformed other methods because the TD-CNN learned more temporal and spatial information than others by capturing the correlation within all hidden states of the LSTM. We also investigated the influence of the value of the time step and concluded that our model performed best when the timestep was 16. Figure 3.4(a) shows the receiver operating characteristic (ROC) curve when timestep=16. The micro-average AUC for our proposed model is 0.99, and the AUC for "severe" is smaller than any other, which indicates it is more challenging for the model to distinguish severe depression from the other levels correctly. This is likely attributed to the absence of severely ill patients in our dataset. Figure 3.4(b) is the ROC when the time step is

32. The micro-average AUC for this model is 0.94. The performance of the model with 32-timesteps was worse than that of the model with 16-timesteps. This is likely due to the negative correlation between the signal-noise ratio of the input sequence and the length of the sequence. A longer input sequence contains more information to assess the emotional state, but as the sequence length grows, the increasing noise cannot be ignored and the bias of the model rises due to the noise. Another factor is the limitation of the memorization capability of LSTM. The longer the input sequence is, the more difficult it is for LSTM to memorize earlier information when processing the end of the sequence because the depth of the LSTM network is proportional to the timestep. Given a long sequence, the information cannot smoothly flow through the network, which results in diminished performance. The confusion matrix of the 32-time step model is illustrated in Figure 3.5(b), which shows the performance of the model on the test partition of the DAIC-WOZ dataset. Comparing the models with different time steps, Figure 3.5(a) shows the confusion matrix of the model with 16 timesteps, while Figure 3.5(c) shows the confusion matrices of the model with 64 timesteps. Different timestep means the different sizes of the test set. To eliminate the influence of the size of the test set, we normalized the confusion matrix along each row. In terms of the normalized confusion matrix, the model with 16 timesteps performed the best, but from the entries on the second row of Figure 3.5(c), the model with 64 timesteps was less likely to classify the mild patients incorrectly. The contribution of the model with a longer time step in the depression prediction should be further investigated to find the cut-off value of the time step that optimizes the trade-off between the computation cost (larger time step means more computation) and the misdiagnosed rate.

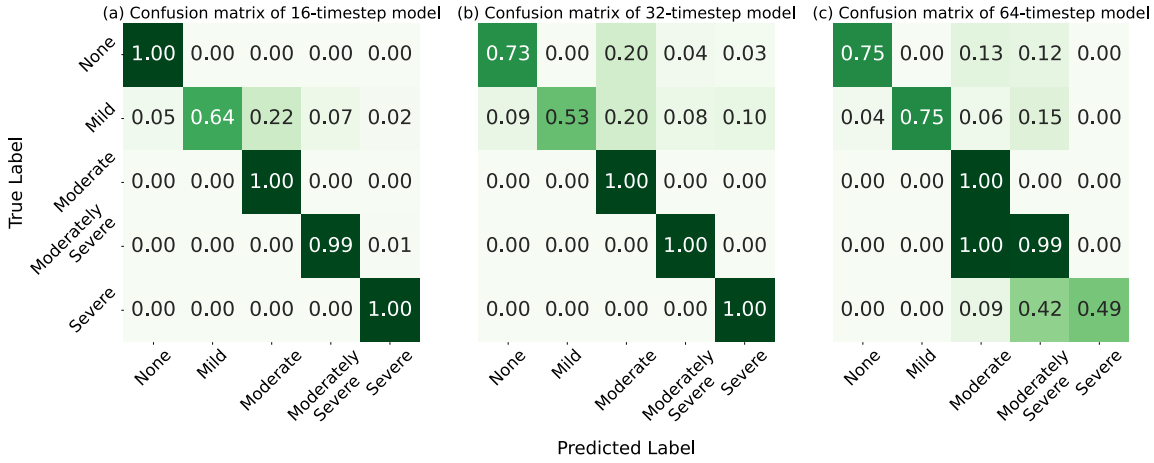

Figure 3.5: (a) Confusion matrix of 16-timestep model on DAIC-WOZ (b) Confusion matrix of 32-timestep model on DAIC-WOZ (c) Confusion matrix of 64-timestep model on DAIC-WOZ

Table 3.5: A Comparative Study of Different Proposed Audio Models

Models	Experimental Settings	Accuracy	F1
LSTM + FC	TS=16, HU=73, LHU=(128,64,5), Adam	0.5674 ± 0.0034	0.5650 ± 0.0042
Bi-LSTM + FC	TS=16, HU=73, LHU=(128,64,5), Adam	0.8717 ± 0.0013	0.8818 ± 0.0013
LSTM + TD-CNN	TS=16, HU=73, #TCNNB=5, #KRNL=(64,64,64,128,256), KS=(3,3,3,3,9), Adam	0.8698 ± 0.0897	0.8609 ± 0.0988
Bi-LSTM + TD-CNN	**TS=16, HU=73, #TCNNB=5, #KRNL=(64,64,64,128,256), KS=(3,3,3,3,9), Adam**	$\mathbf{0.9871 \pm 0.0009}$	$\mathbf{0.9870 \pm 0.0009}$

#TCNNB: Number of TD-CNN blocks #KRNL: Number of conv kernels in each TD-CNN block #KS: Kernel size

3.3.3 Results of the Text Modality

The Effect of Stop Words and Bidirectional Layer

In this experiment, we used NLTK to remove the stop words in English transcripts. Apart from the stop words, the other factor is the choice between LSTM and Bi-LSTM models. Compared with the unidirectional LSTM model, the bidirectional model converges faster, and the validation accuracy is higher. The following experiment demonstrates several advantages of the Bi-LSTM model over the traditional LSTM model on the depression level classification task. Four models were trained with the different configurations presented in Table 3.6. The test set for evaluation is balanced by oversampling the minority class.

Table 3.6: A Comparative Study of the Proposed Text Models

Models	Experimental Settings	Accuracy	F1	Micro-average AUC
LSTM + FC	TS=64, HU=100, LHU=(256,128,5), Adam, Stopwords	0.9091	0.9094	0.9738
LSTM + FC	TS=64, HU=100, LHU=(256,128,5), Adam, No stopwords	0.9792	0.9754	0.9897
Bi-LSTM + FC	TS=64, HU=100, LHU=(256,128,5), Adam, Stopwords	0.9617	0.9610	0.9908
Bi-LSTM + FC	**TS=64, HU=100, LHU=(256,128,5), Adam, No stopwords**	**0.9685**	**0.9709**	**0.9925**

TS: Timestep; HU: #Hidden units in LSTM; LHU: #Hidden units in linear layers

From Table 3.6, we concluded that if the type of the LSTM was fixed (i.e., the two text models both consist of LSTM or Bi-LSTM network), the performance of the model without stop words was better. If the stop words were kept, the Bi-LSTM model still outperformed the traditional one. This result was in line with our expectation that Bi-LSTM was better in text classification because it learned more contextual information with the combination of the forward and backward networks.

The Effect of Window Size

Window size is another factor that influences the performance of the model. Intuitively, the longer the window, the more information it contains about the mental state of the subjects, which means our model can assess their emotions more accurately. However, if the window is too long, while making an inference, the impact of the noise cannot be ignored, which leads to significant performance degradation. Moreover, the memorization capability of LSTM is limited, which means the longer the sequence is, the more challenging for the LSTM to memorize and extract useful information. To demonstrate the relationship between the performance and the window size, we conducted experiments by changing the window size. As shown in Table 3.7, when the window size started to increase, the metrics increased firstly but began to decrease after the window size is greater than 64. This was in line with our expectation, the classifier gained a lot of information due to a larger window but started to degrade as the result of the noise in the large window and the reduced performance of LSTM. We concluded that the window size should be appropriately set to train the model with the best performance, in our experiment, the best window size is 64.

Table 3.7: A Comparative Study of the Text Model with Different Window Size

Window Size	Accuracy	Precision	Recall	F1 Score
16	0.8254	0.8318	0.8340	0.8141
32	0.8256	0.8371	0.8465	0.8260
64	**0.8778**	**0.8779**	**0.8782**	**0.8705**
128	0.8409	0.8599	0.8430	0.8304

Table 3.8: A Comparative Study of Our Proposed Patient-Level Methods and the State of the Art

Model	Experimental Settings	Accuracy	F1	Sensitivity	Specificity
UniLSTM as encoder	WIN=16, Stride=64	0.8604	0.8579	0.9844	0.8182
	WIN=32, Stride=64	0.9209	0.9188	0.9647	0.9777
	WIN=64, Stride=64	0.8674	0.8682	0.9705	0.9888
BiLSTM as encoder	WIN=16, Stride=64	0.9488	0.9500	0.9735	0.9444
	WIN=32, Stride=64	0.9186	0.9191	0.9852	0.9700
	WIN=64, Stride=64	0.8535	0.8546	0.9647	0.8778
BiLSTM as encoder	WIN=16, Stride=64, attention	0.8419	0.8427	0.9735	0.8222
	WIN=32, Stride=64, attention	**0.9581**	**0.9580**	**0.9824**	**1.0000**
	WIN=64, Stride=64, attention	0.9093	0.9086	0.9706	0.9889
UniLSTM as encoder	WIN=16, Stride=64, attention (aligning&fusion)	0.8977	0.8973	0.9559	0.9889
	WIN=32, Stride=64, attention (aligning&fusion)	0.9326	0.9315	0.9735	0.9889
	WIN=64, Stride=64, attention (aligning&fusion)	0.8581	0.8615	0.9412	0.8889
BiLSTM as encoder	WIN=16, Stride=64, attention (aligning&fusion)	0.8491	0.8439	0.9353	0.9000
	WIN=32, Stride=64, attention (aligning&fusion)	0.9047	0.9103	0.9941	0.9000
	WIN=64, Stride=64, attention (aligning&fusion)	0.6279	0.6560	0.7500	1.0000
Unimodality text model	WIN=16	*	BLSTM: 0.7929 ULSTM:0.8096	*	*
	WIN=32	*	BLSTM: 0.7964 ULSTM:0.7619	*	*
	WIN=64	*	**BLSTM: 0.9245 ULSTM:0.9058**	*	*
	WIN=128	*	BLSTM: 0.8266 ULSTM:0.7148	*	*
Unimodality audio model	BLSTM + FC	*	0.8819	*	*
	ULSTM + FC	*	0.7604	*	*
	BLSTM + TCNN	*	**0.9074**	*	*
	ULSTM + TCNN	*	0.8443	*	*
[206]	End to end convolutional neural network	0.7464	0.7750	0.74	0.8
[200]	Combination of LSTM and CNN	*	0.77	0.83	*
[207]	Hierarchical context-aware graph attention model	*	0.92	0.92	*

3.3.4 Results of the Fused Model

In this experiment, the audio and text models were jointly optimized so that we could verify whether our methods were still effective under multimodality configuration. We proposed three varieties of fusion models and merged these segment-wise predictions through major voting to obtain the patient-level prediction. The configuration details of those fused models were described in Section 3.2.4. The metrics of each fusion model on the test partition were covered in Table 3.8. When experimenting with models made up of unidirectional LSTM, without an attention mechanism, the model with a window size of 32 performed better than others when classifying for a multi-class outcome in terms of the accuracy on the test set ($accuracy = 0.9209$). Theoretically, the models with Bi-LSTM should be better than a uni-LSTM one, however, with all other configurations fixed, except the Bi-LSTM model with a window size of 16, other Bi-LSTM models did not show significant improvement over the uni-LSTM one. Nevertheless, once the attention mechanism was introduced, the performance was boosted and the F_1 increased compared to the model without an attention mechanism, except the Bi-LSTM model with an attention mechanism and window size of 16. As we reported in the methodology section, the attention mechanism could be introduced during the multimodal feature aligning phase as well as the multimodality fusion phase. The attention mechanism during the fusion process weighed each modality and made it possible for the model to determine the contribution of each modality. From Table 3.8, we concluded that the highest sensitivity of 0.9941 was achieved by the model comprised of Bi-LSTM and two attention layers, with a window size of 32. Given that we expected to train an early-stage depression screening tool, we preferred higher sensitivity so that we would not miss those potential depression patients. The model with two attention layers led to results that outperformed the state of the art, Niu et al., by 8% in terms of sensitivity. In comparison with Alhanai et al., who adopted a similar method made up of CNN and LSTM,

our proposed method was better by 17% in terms of sensitivity. This is not conclusive since the dataset for evaluation in their article was slightly different from ours. By conducting a student t-test between the F_1 of the best patient-level audio model with the result of p-value = 0.0099 (<0.01), cohen's d = 2.1258, as well as patient-level text model with the result of p-value = 0.0246, (<0.05), cohen's d = 1.7452, we could conclude that multimodality models statistically significant outperformed single modality models.

3.4 Conclusion

In this chapter, a multimodality approach for automated depression detection was presented. Firstly, we performed the statistical test to investigate the difference between the audio and text features of severe and healthy subjects. We proved the pattern of severe depression patients was different from that of the healthy. Therefore, the audio feature sequence carried information that could be used to predict depression severity. Secondly, models that considered audio and text features individually were trained and evaluated at the patient-independent level. These unimodality models then acted as feature extractors and output features were combined by an audio-text fused model. For the audio modality, at the patient-independent level, the model comprised of single-layer Bi-LSTM and five stacked TD-CNN blocks achieved the best sequence level F_1 score of 0.9870 and patient-level F_1 score of 0.9074 with the test set. This result indicates that the Bi-LSTM provides a more reliable representation, from which the automated depression detection model could benefit. Additionally, we evaluated the patient-independent audio models with different timesteps with the Area Under Curve (AUC) metric. We concluded that the 16-timestep model performed best and the micro-average AUC was higher than any other model. However, the 64-timestep model showed its strength in detecting the audio feature sequence from the mild patient, which met our expectation that the model should be able to

distinguish mild patients so that clinical interference can be conducted in the early stage. Overall, the 16-timestep model outperformed the 32-timestep and 64-timestep models, which could be attributed to the relatively low signal-noise ratio of the shorter input sequences and the memorization limit of the LSTM. The new understanding assisted in our model selection and hyper-parameter configuration when we deployed this method in clinical settings. These findings provided the following insight for future research, our proposed unimodality model was patient-independent, and the prediction was based on a period of audio/text features. Therefore, compared with other models, our proposed model did not have limitations to the length of the interview audio or transcript, which made it possible for people to monitor their mental state in daily use.

Moreover, for the text modality, the model consisting of Bi-LSTM and three fully connected layers achieved the best sequence level F_1 score of 0.9709 and patient-level F_1 score of 0.9245 on the test set. We conducted experiments to investigate the influence of the text model hyper-parameters, such as window size and stop words. We found the best window size is 64. In our experiment, we investigated the effect of stop words, the result indicated the text model performs better if the stop words were removed in advance. Currently, our patient-level prediction was carried out by a major voting algorithm, which yielded a patient-level depression prediction model with satisfying performance. Our proposed multimodal method achieved the highest F_1 of 0.9580 on the patient-level depression detection task, which showed a significant improvement over the previous state-of-the-art. In the future, a study on how to represent the audio/text features during the whole interview should be carried out so that the model could make patient-level predictions based on a digest of text and audio features.

Chapter 4

Analysis for Automated Clinical Depression Diagnosis in a Chinese Corpus

4.1 Introduction

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 datasets have investigated music-induced [34], video-induced [29, 34] and mixed emotion induction methods [35]. 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 these issues, there is a need for interactive multimodal datasets for the study of depression, collected through interviews conducted in a clinical setting between patients and physicians.

In this scenario, patients 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 evaluating depression risk. To ensure the col-

lection 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) [185]. The audio features, such as formant frequency F_0 and Normalized Amplitude Quotient (NAQ), were then extracted using the COVAREP toolbox, and the interview recordings were transcribed verbatim using an audio transcription application programming interface (API) developed by iFlyTek for subsequent analysis [54]. 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. This study introduces a new corpus comprising interviews conducted with clinically depressed patients, gathered from a psychiatric hospital, containing 113 recordings with 52 healthy and 61 depressive patients. 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 is also investigated and illustrated. To the best of our knowledge, this is the first study to compile a corpus of clinical interviews focused on depression in Chinese, and to subsequently train machine learning models to identify depression patients.

In this chapter, 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 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.

4.2 Related Work

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

4.2.1 Data Collection Methods

Data collection methods play a crucial role in impacting model performance. Researchers must consider an appropriate context in which subjects' responses are observed. So far, 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 corpus [208, 209]; spontaneous behaviour - interviewees naturally interact with interviewers or machines, for example, chatting with a chatbot [144, 210]. 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.

Social Network Platform

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. On the flip side, many researchers have had to perform feature extraction from scratch and design application-specific machine learning strategies due to data scarcity.

Collecting data from online forums can also significantly reduce the difficulty of obtaining sufficient data from healthy individuals. The healthy control group can be sampled from other online communities unrelated to depression [73, 115, 211]. 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 clinic attendance due to fear of mental disorder-related stigma may also avoid discussing depression online.

Interview Under Controlled Conditions

Many researchers are turning to 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 [212], which includes audio and video recordings of 150 participants and the Affectiva-MIT Facial expression dataset (AM-FED) [213], which consists of labelled spontaneous facial recordings collected over the internet, including 242 video recordings and labels of the presence of 10 symmetrical and 4 asymmetrical action units (AU), head movements, smiles, feature tracker confidence, as well as biological sex and facial landmarks.

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 [144]. This 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. Suppose 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, which may not contain enough information to assess their emotional state accurately.

4.2.2 Depression Assessment Instrument

Table 4.1: A Comparative Study of Our Proposed Dataset and Datasets Employed by the Reviewed Studies for Depression Detection

Dataset	Population/ Healthy v.s depressive	Collection protocol	Language	Label	Criteria	Research purposes	Video resolution	Modality	Controlled condition
[214]	770	Social media	English	Self-report	-	Detection	-	Text	NA
[209]	753	Social media	English	Self-report		Detection	-	Text	NA
[215]	49	Interpersonal & Virtual agent	English	Clinical assessment	DSM-IV, HAMD>15	Detection	640x480	Visual & Audio	Yes
[150]	130 (70/60)	Interpersonal & Virtual agent	English	Clinical assessment	DSM-IV, HAMD>15	Detection	800x600	-	Yes
[144]	189 (132/57)	Interpersonal & Virtual agent	English	Self-report	PHQ-8>10	Detection & Severity	-	Visual & Audio & Text	Yes
[216]	58	Virtual agent	English	Self-report	-	Severity	-	Visual & Audio	Yes
[124]	887	Social media	English	Self-report		Detection	-	Text	NA
[217]	8 (4/4)	Interpersonal	English	-	-	Detection	-	-	Yes
[218]	30 (15/15)	Interpersonal	English	Clinical assessment	-	Detection	-	-	Yes
[219]	26 (13/13)	Interpersonal & Virtual agent	Chinese	Clinical assessment	HAMD>15	Detection	640x480	-	Yes
[59]	7 (0/7)	Interpersonal	English	Clinical assessment	HAMD>15	Recovery	-	-	No
[220]	78 (26/52)	Interpersonal	Chinese	Clinical assessment	HAMD>17 or PHQ-9≥9	Detection & Severity	-	Visual & Audio & Text	No
Ours	**113 (52/61)**	**Interpersonal**	**Chinese**	**Clinical assessment**	-	**Detection & Severity**	-	**Audio & text**	**No**

4.2.3 The Existing Corpora

We reviewed previous articles that reported on dyadic interview recordings that were annotated based on clinician and outpatient interactions. We found that about half of these datasets were labelled with a self-reported depression rating scale, recorded in controlled conditions, and produced in English. The controlled condition refers to the standardized task and procedures that we use during the interviews. Specifically, in

previous studies, researchers asked the interviewees to perform tasks such as reading a fixed paragraph, sustained vowels, and memory recalling, which allows them to control for variability in responses and simplify the problem. In terms of the differences between our dataset and others, we believe that our approach allows us to collect more natural responses from participants. While previous studies have used structured interviews, our interviews allow for more flexibility in the conversation, as participants can choose to continue or change the topic as they see fit. This approach can yield more spontaneous and authentic responses from the participants, which is important for accurately diagnosing depression. The prevalence of one language in these datasets limits their usability for cross-cultural studies of depression. Table 4.1 compares existing data from social networks and clinical interviews.

4.3 Methods and Procedure

4.3.1 Data Collection

The main goal of this study is 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 [221, 222]. In addition, the study aims to examine the 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.

Participants

113 participants were recruited for a psychology study, with informed consent and a range of ages from 15 to 65, who were native Mandarin speakers with at least primary education. To ensure that our findings are applicable to a broader population, we took care to select a representative sample of individuals with depression. We also

took into account the recommendations from clinicians and excluded individuals with a history of antidepressant medication or mental disorders from our study. Participants who were diagnosed with depression had no other mental or medical conditions. 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 more reflective of real-world exposure to noise pollution. If the model is trained on noise-cancelling data, its performance may not be as good in real-world settings where noise cancellation is not available. We ensured the standardization of our data collection process in several ways. Firstly, we employed experienced physicians to conduct all the interviews. We also utilized the MADRS questionnaire, a reliable and well-established tool with high inter-rater consistency and reliability. By using such questionnaires and qualified clinicians, we could ensure that the data collection process was standardized. Apart from using a reliable questionnaire and experienced physicians, we have established standardized protocols for conducting interviews, recording data, and managing data quality. Before starting the full-scale data collection, we conducted pilot testing, like randomly picking some outpatients to conduct depression interview to ensure the data collection process is feasible, reliable and standardized. In addition, we developed a data management plan that outlines procedures for storing, protecting and sharing data to ensure that data quality and privacy are maintained throughout the data collection process. Clinicians were not aware of the condition of the examined subject in advance. Interviews

were conducted in Mandarin, lasting 5-10 minutes, and based on the MADRS questionnaire [185]; 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).

Procedure

After obtaining verbal consent and a signed form for recording, the clinician administered the MADRS questionnaire in Chinese to the participant. 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 [185]. The questionnaire focused on the ten critical symptoms in Table 4.2.

Table 4.2: The Questionnaire Used During Interview

Symptom	Question
Apparent sadness	Not Applicable
Reported sadness	How is everything going?
Inner tension	Have you ever been feeling nervous and scared for no reason?
Reduced sleep	How do you sleep recently?
Reduced appetite	How do you eat recently?
Concentration difficulties	Can you stay focused?
Lassitude	Do you feel like you don't want to do anything?
Inability to feel	Do you feel that everything has nothing to do with you?
Pessimistic thoughts	Do you feel inferior or self-blaming?
Suicidal thoughts	Have you ever thought of self-harm or suicide?

During the experiment, participants were asked questions by a clinician about their mental health. The questions may not have been in the exact order of the

MADRS questionnaire, and the clinician may have asked additional questions for more information based on their 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 adapt 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 physicians 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.

Dataset Statistics

In our study, we interviewed 113 participants, 52 of whom were healthy, and 61 were patients. The interview audios are an average of 364.40 seconds in length (st. dev = 257.66 seconds). For the control group, the audios have 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 experimental group, the audios 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 4.3. Before further analysis, building a balanced dataset by random sampling is important. For binary depression detection, the positive and negative samples should be approximately equal. For the multiclass depression level prediction, the distribution of severity levels should be balanced. In our dataset, 52 and 61 participants were in the control and experiment groups. Figure 4.1(a) shows the distribution of the depression levels in our dataset. Figure 4.1(b) and Figure 4.1(d) illustrate the distribution of the audio duration in healthy and depressive groups. The average audio duration for the depressive population was significantly longer than that of the healthy ($p < 0.01$, cohen's d = 1.94). The distribution of the utterance length was given in Figure 4.1(c) and Figure 4.1(e); as opposed to the distribution of audio duration, the number of words in a sentence for the control and experiment groups was

not significantly different ($p > 0.05$, cohen's $d = 0.06$). To identify patterns between the healthy and depressive groups, we generated word clouds and showed frequently used words in a larger font in Figure 4.1(f) and Figure 4.1(g). The word cloud of the depressive reveals that depressive 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.

Table 4.3: Summary of Dataset Characteristic

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

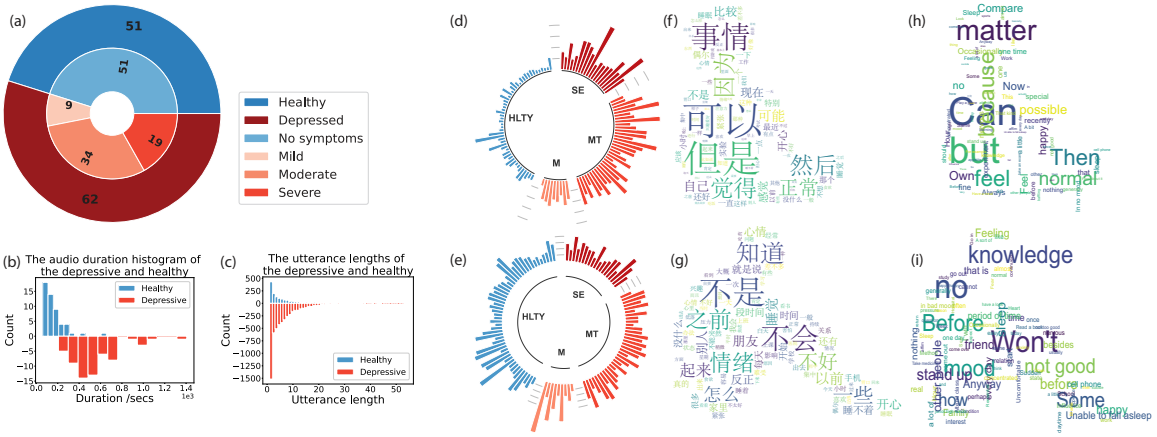

Figure 4.1: The proposed dataset contains 113 individuals, (a) 51 of whom are healthy and 62 of whom are patients with depression. Of the patients with depression, 9 have mild depression, 34 have moderate depression, and 19 have severe depression. (b) The distribution of audio duration between the healthy and depressive. (c) The distribution of utterance length between the healthy and depressive. (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). (h), (i) The word cloud in English. More negative words, such as "difficult to fall asleep", "bad mood", etc., are in the word cloud below.

4.3.2 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.

Audio Transcription

The iFlyTek Audio transcribes 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.

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" 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 the "word 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 research assistants verified the transcription against the audio recordings, 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."

Audio Feature Extraction

We used the collaborative voice analysis repository (COVAREP) toolkit to capture the frame-level acoustic features [54]. COVAREP is an open-source feature extraction toolkit commonly used in depression classification studies. We segment the raw interview audio recordings using the Cooperative Voice Analysis Repository (COVAREP) toolkit, which allows us to extract audio features at a rate of 100Hz. To realize this, we divide the raw recording into blocks, with each block lasting 10 milliseconds, which is a common practice in speech processing [223–226]. 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 F0, voiced/unvoiced (VUV) decision, normalized amplitude quotient (NAQ), quasi-open quotient (QOQ), H1-H2, peak-slope (PSP), modulation depth quotient (MDQ), relative amplitude quotient (Rd), creaky voice detection, Mel-Cepstral coefficients (MCEPs), Harmonic Model + Phase Distortion (HMPD) features. By using a 10-millisecond block size and a sampling rate of 100Hz, we believe that we are able to capture the relevant acoustic information at a reasonable computational cost. Detailed descriptions of each audio feature can be found in Table 4.4.

Table 4.4: COVAREP Spectral and Cepstral Feature Set

Voicing based	**Group**
F0	Spectral
VUV	-
PSP	Spectral
Glottal source-based	**Group**
Normalized amplitude quotient(NAQ)	Spectral
QOQ	Spectral
H1, H2	Spectral
Parabolic Spectral Parameter (PSP)	Spectral
Spectral envelope-based	**Group**
Mel-cepstral coefficients ($MCEP_0$-$MCEP_{25}$)	Cepstral
Harmonic model and phase distortion mean ($HMPDM_0$-$HMPDM_{25}$)	Cepstral
Harmonic model and phase distortion deviation ($HMPDD_0$-$HMPDD_{13}$)	Cepstral
Wavelet-based	**Group**
Maxima Dispersion Quotient (MDQ)	Spectral
Peak slope	Spectral

4.3.3 Baseline Audio Models

We used COVAREP to extract audio features from each participant at a rate of 10 milliseconds. The interviews lasted between 5 to 10 minutes, resulting in variable numbers 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.

To make the audio feature frames fixed-length, we computed the histogram of each audio feature to obtain 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 [227]. 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.

4.3.4 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 the 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, bi-gram, Term Frequency-Inverse Document Frequency (TF-IDF), 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, on the other hand, 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. Term Frequency-Inverse Document Frequency (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 $x = (x_1, x_2, \ldots, x_n), x \in R^{s \times n}$, 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 and level based on each subject's feature vector., i.e.

$$P\left(C = c_i \mid X = x_i\right) = \frac{P\left(X = x_i \mid C = c_i\right) P\left(C = c_i\right)}{P\left(X = x_i\right)} \tag{4.1}$$

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

$$P\left(X = x_i \mid C = c_i\right) P\left(C = c_i\right) \tag{4.2}$$

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

$$h^*\left(x\right) = \arg\max_i P\left(X = x_i \mid C = c_i\right) P\left(C = c_i\right) \tag{4.3}$$

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.

$$f_i(x) = \Pi_{j=1}^{n} P(X_j = x_j \mid C = c_i) P(C = c_i) \tag{4.4}$$

The set of conditional probabilities in Equation 4.4 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 $p(x|y, c_i)$, which is given by:

$$P(x \mid y, c_i) = \frac{P(x, y, c_i) + \alpha \cdot P(x)}{P(y, c_i) + \alpha} \tag{4.5}$$

4.4 Results and Discussion

4.4.1 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 4.1), most clinical datasets comprise 100 to 200 data points, due to the high cost of data collection.

Experimental Setting

This study focused on detecting depression and predicting the severity of depression. 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 had 10 healthy individuals and 13 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, with 41 healthy individuals, 7 with mild depression, 27 with moderate depression, and 15 with severe depression in the training set, while the test set had 10 healthy individuals, 2 with mild depression, 7 with moderate depression, and 4 with severe depression.

To maintain balance among minority classes, we oversampled 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 [228] on the proposed dataset. For the deep learning-based experiments, we labelled each frame of audio and text features by 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 in [228].

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.

Multi-classification Text Model

To investigate whether the severity of depression is 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 4.4.1, the training set had 41 healthy individuals, 7 with mild depression, 27 with moderate depression, and 15 with severe depression, while the test set had 10 healthy individuals, 2 with mild depression, 7 with moderate depression, and 4 with

Table 4.5: Cross Validation and Testing Result of the Text Depression Detection Model

Text modality (Binary classification, nested cross-validation, testing)											
Model	State	Precision		Recall		F1 score		Accuracy		Cohen Kappa	
Uni-gram	Healthy	0.94 ± 0.07	1	0.73 ± 0.09	0.80	0.82 ± 0.06	0.89	**0.82 ± 0.03**	**0.91**	**0.70 ± 0.02**	**0.82**
	Depressive	0.81 ± 0.05	0.87	0.96 ± 0.07	1	0.88 ± 0.05	0.93				
	avg.	**0.87 ± 0.06**	**0.92**	**0.86 ± 0.05**	**0.91**	**0.85 ± 0.03**	**0.91**				
Bi-gram	Healthy	0.96 ± 0.09	1	0.56 ± 0.02	0.70	0.71 ± 0.03	0.82	0.79 ± 0.02	0.87	0.56 ± 0.06	0.73
	Depressive	0.73 ± 0.04	0.81	0.98 ± 0.04	1	0.83 ± 0.06	0.90				
	avg.	0.83 ± 0.09	0.89	0.79 ± 0.05	0.87	0.78 ± 0.09	0.86				
TF-IDF	Healthy	1	1	0.59 ± 0.12	0.70	0.74 ± 0.11	0.82	0.81 ± 0.05	0.87	0.61 ± 0.05	0.73
	Depressive	0.74 ± 0.08	0.81	1	1	0.85 ± 0.06	0.90				
	avg.	0.86 ± 0.07	0.89	0.81 ± 0.03	0.87	0.80 ± 0.06	0.86				
[228]	Healthy	-	1	-	0.96	-	0.98	-	**0.97**	**-**	**0.86**
	Depressive	-	0.95	-	1	-	0.97				
	avg.	-	**0.98**	-	**0.97**	-	**0.97**				

[a] Underline indicates results on the test set.

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 F_1 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 F_1 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 α in the range of 10^k where k was set to 0, 1, 2, or 3. The best value for α 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 4.6.

Table 4.6: Cross Validation and Testing Result of the Text Depression Level Prediction Model

Text modality (depression level prediction, nested cross-validation, testing)											
Model	State	Precision		Recall		F1 score		Accuracy		Cohen kappa	
Uni-gram	None & Mild	0.91 ± 0.04	0.73	0.76 ± 0.08	0.80	0.83 ± 0.07	0.76	0.58 ± 0.06	0.61	0.36 ± 0.02	0.38
	Moderate & Severe	0.43 ± 0.06	0.50	0.74 ± 0.08	0.86	0.54 ± 0.08	0.63				
	avg.	0.57 ± 0.07	0.47	0.58 ± 0.06	0.61	0.55 ± 0.09	0.52				
Bi-gram	None & Mild	0.89 ± 0.08	1	0.80 ± 0.05	0.80	0.85 ± 0.04	0.89	**0.62 ± 0.04**	**0.65**	**0.41 ± 0.02**	**0.47**
	Moderate & Severe	0.46 ± 0.05	0.47	0.81 ± 0.07	1	0.59 ± 0.03	0.64				
	avg.	**0.59 ± 0.03**	**0.58**	**0.62 ± 0.05**	**0.65**	**0.58 ± 0.04**	**0.58**				
TF-IDF	None & Mild	0.79 ± 0.05	0.80	0.90 ± 0.03	0.80	0.84 ± 0.04	0.80	0.66 ± 0.05	0.57	0.44 ± 0.03	0.32
	Moderate & Severe	0.51 ± 0.04	0.38	0.81 ± 0.05	0.71	0.63 ± 0.05	0.50				
	avg.	0.51 ± 0.02	0.46	0.66 ± 0.04	0.57	0.57 ± 0.02	0.50				
[228]	None & Mild	-	1	-	0.96	-	0.98	-	**0.85**	-	**0.87**
	Moderate & Severe	-	0.64	-	1	-	0.78				
	avg.	-	**0.77**	-	**0.85**	-	**0.80**				

[a] Underline indicates results on the test set.

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 attractive properties: firstly, it is an ensemble learning method, which decreases the bias of the model; it is a tree-based model with high interpretability, determining the feature's importance in making an inference. These models offer a good trade-off between computation cost and accuracy. Tree-based boosting algorithm methods solve many machine learning problems efficiently and accurately, making it a good candidate for providing baseline results within our dataset.

To train our XGBoost classifiers, we created a separate model for each audio feature. However, we excluded certain features, including HMPDM_0 to HMPDM_3, since they remained constant throughout the interview. To make 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 the other metrics can be found in Table 4.7.

Table 4.7: Cross Validation and Testing Result of the Audio Depression Detection Model

Audio modality (Binary-classification, nested cross-validation, testing)											
Model	State	Precision		Recall		F1 score		Accuracy		Cohen Kappa	
XGBoost	Healthy	0.83 ± 0.04	0.89	0.73 ± 0.05	0.80	0.78 ± 0.02	0.84	**0.81 ± 0.04**	**0.87**	**0.62 ± 0.06**	**0.73**
	Depressive	0.80 ± 0.04	0.86	0.88 ± 0.04	0.92	0.83 ± 0.05	0.89				
	avg.	**0.81 ± 0.07**	**0.87**	**0.81 ± 0.05**	**0.87**	**0.81 ± 0.04**	**0.87**				
Decision Tree	Healthy	0.75 ± 0.04	0.78	0.66 ± 0.03	0.70	0.70 ± 0.05	0.74	0.74 ± 0.05	0.78	0.48 ± 0.02	0.55
	Depressive	0.74 ± 0.03	0.79	0.82 ± 0.04	0.85	0.78 ± 0.05	0.81				
	avg.	0.74 ± 0.03	0.78	0.74 ± 0.02	0.78	0.74 ± 0.04	0.78				
[228]	Healthy	-	0.78	-	0.96	-	0.98	-	**0.97**	-	**0.95**
	Depressive	-	0.95	-	1	-	0.97				
	avg.	-	**0.98**	-	**0.97**	-	**0.97**				

[a] Underline indicates results on the test set.

Multi-classification Audio Model

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 we then used in our analysis. We trained and evaluated a set of models on the training and validation set, testing different parameters to optimize performance. We selected the model with the highest F_1 score for our analysis. This model achieved an F_1 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 4.8.

Table 4.8: Cross Validation and Testing Result of the Audio Depression Level Prediction Model

Audio modality (depression level prediction, nested cross-validation, testing)											
Model	State	Precision		Recall		F1 score		Accuracy		Cohen Kappa	
XGBoost	None & Mild	0.72 ± 0.03	0.62	1	1	0.84 ± 0.05	0.77				
	Moderate & Severe	0.42 ± 0.02	0.71	0.52 ± 0.03	0.71	0.47 ± 0.02	0.71	$\mathbf{0.61 \pm 0.02}$	**0.65**	$\mathbf{0.35 \pm 0.02}$	**0.43**
	avg.	$\mathbf{0.45 \pm 0.02}$	**0.49**	$\mathbf{0.61 \pm 0.03}$	**0.65**	$\mathbf{0.52 \pm 0.04}$	**0.55**				
Decision Tree	None & Mild	0.72 ± 0.03	0.62	1	1	0.84 ± 0.04	0.77				
	Moderate & Severe	0.48 ± 0.02	0.57	0.59 ± 0.03	0.57	0.53 ± 0.01	0.57	0.63 ± 0.04	0.61	0.39 ± 0.01	0.35
	avg.	0.47 ± 0.02	0.45	0.63 ± 0.04	0.61	0.54 ± 0.03	0.51				
[228]	None & Mild	-	1	-	0.95	-	0.97				
	Moderate & Severe	-	0.60	-	1	-	0.75	-	**0.81**	-	**0.95**
	avg.	-	**0.88**	-	**0.81**	-	**0.76**				

[a] Underline indicates results on the test set.

Multimodality Baseline Models

To enhance the 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 4.5 and Table 4.7, 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 4.6 and Table 4.8. During cross-validation and on the test set, our multimodality depression detection model (accuracy=0.86, see Table 4.9) and multimodality depression-level classification model (accuracy=0.63, see Table 4.10) produced fewer errors than the acoustic-only and semantic-only models.

4.4.2 Discussion on Audio Features Statistics

Our dataset analyzed recordings from 113 clinically supervised participants, resulting in two different comparisons: inter-condition and intra-condition comparisons. Inter-condition comparison evaluates the differences in audio features between healthy and

Table 4.9: Cross Validation and Testing Result of the Multimodality Depression Detection Model

Fused model (Binary classification, nested cross-validation, testing)											
Model	State	Precision		Recall		F1 score		Accuracy		Cohen Kappa	
Uni-gram + XGBoost	Healthy	0.94 ± 0.02	1	0.73 ± 0.04	0.80	0.82 ± 0.03	0.89	**0.86 ± 0.04**	**0.91**	**0.72 ± 0.04**	**0.78**
	Depressive	0.81 ± 0.02	0.87	0.96 ± 0.03	1	0.88 ± 0.03	0.93				
	avg.	**0.87 ± 0.02**	**0.92**	**0.86 ± 0.03**	**0.91**	**0.85 ± 0.01**	**0.91**				
Bi-gram + XGBoost	Healthy	0.96 ± 0.02	1	0.56 ± 0.03	0.70	0.71 ± 0.02	0.82	0.79 ± 0.02	0.87	0.68 ± 0.06	0.72
	Depressive	0.73 ± 0.02	0.81	0.98 ± 0.01	1	0.83 ± 0.04	0.90				
	avg.	0.83 ± 0.03	0.89	0.79 ± 0.02	0.87	0.78 ± 0.04	0.86				
TF-IDF + XGBoost	Healthy	1	1	0.59 ± 0.02	0.70	0.74 ± 0.02	0.82	0.81 ± 0.04	0.87	0.65 ± 0.04	0.69
	Depressive	0.74 ± 0.03	0.81	1	1	0.85 ± 0.04	0.90				
	avg.	0.86 ± 0.03	0.89	0.81 ± 0.05	0.87	0.80 ± 0.02	0.86				

[a] Underline indicates results on the test set.

Table 4.10: Cross Validation and Testing Result of the Multimodality Depression Level Prediction Model

Fused model (depression level prediction, nested cross-validation, testing)											
Model	State	Precision		Recall		F1 score		Accuracy		Cohen Kappa	
Unigram + XGBoost	None & Mild	0.83 ± 0.04	0.91	0.85 ± 0.06	1	0.84 ± 0.03	0.95	**0.63 ± 0.05**	**0.70**	**0.43 ± 0.05**	**0.53**
	Moderate & Severe	0.47 ± 0.02	0.56	0.67 ± 0.04	0.71	0.55 ± 0.03	0.63				
	avg.	**0.59 ± 0.02**	**0.62**	**0.63 ± 0.04**	**0.70**	**0.60 ± 0.01**	**0.65**				
Unigram + Decision Tree	None & Mild	0.85 ± 0.03	1	0.85 ± 0.02	0.80	0.85 ± 0.04	0.89	0.62 ± 0.03	0.57	0.41 ± 0.02	0.35
	Moderate & Severe	0.45 ± 0.04	0.42	0.70 ± 0.03	0.71	0.55 ± 0.04	0.53				
	avg.	0.58 ± 0.01	0.56	0.62 ± 0.03	0.57	0.59 ± 0.03	0.55				

[a] Underline indicates results on the test set.

depressive groups, such as whether participants from the control (healthy) and experimental (depressive) groups differ in vocal fundamental frequency F_0 processed by the top-N elements method. Intra-condition comparisons evaluate the variability of patients' audio features relative to their severity of depression. This second comparison is crucial because it offers a new way to understand if 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 Vocal fundamental frequency (F_0), Mel-Cepstrum Coefficient 0 (MCEP_0), and Harmonic model and phase distortion mean 17 (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 we 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.

Vocal Fundamental Frequency (F_0)

Fundamental frequency 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 fundamental frequency F_0. The median F_0 of the healthy control group is lower compared to participants from the depression group, as shown in Figure 4.2. This is in line with the conclusion reached by Mundt et al. that the healthy control group has a lower F_0 than the depressive experiment group [229]. 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, the F_0 was not a significant audio

feature in male participants.

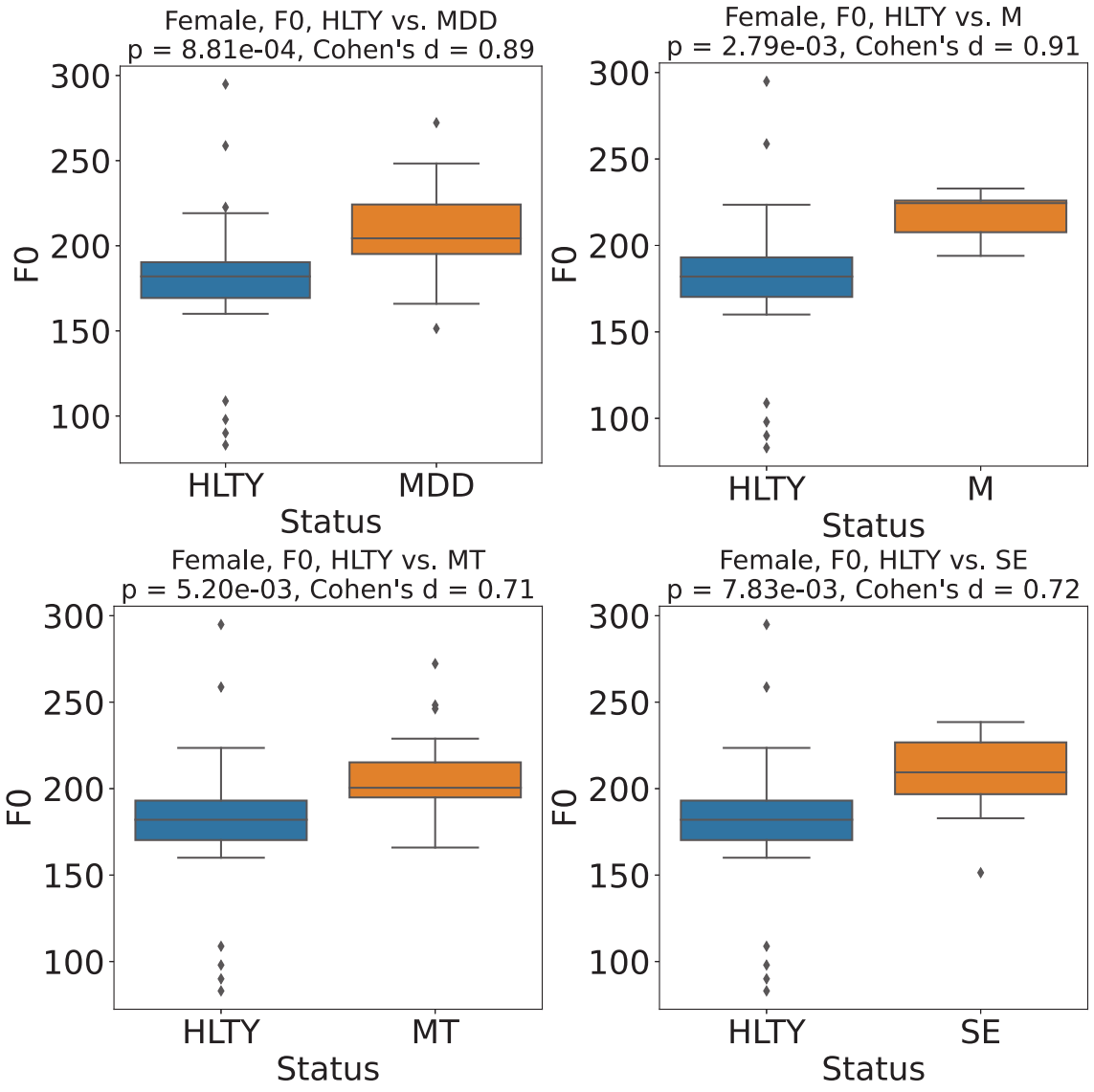

Figure 4.2: The audio feature F_0 of female subjects between healthy and depressive, healthy and mild, healthy and moderate, healthy and severe. (HLTY: Healthy, MDD: Major Depressive Disorder, M: Mild depression, MT: Moderate depression, SE: Severe depression)

Mel-Cepstrum Coefficient (MCEP)

The Mel-Cepstrum coefficient (MCEP) is included in our model as it has been effectively used to characterize speech content [230–232]. 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 some MCEP values significantly different between the healthy and depressive groups. The box plots in Figures 4.3 and 4.4 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 also a reliable indicator of depression severity classification; it was significantly different in the healthy, mild, moderate and severe depression groups, as shown in Figures 4.3 and 4.4. Therefore, $MCEP_0$ may be a gender (biological sex)-independent factor in distinguishing depression presence and severity.

Harmonic Model and Phase Distortion Mean (HMPDM)

Several reports have shown that HMPDMs can be used to predict depression presence and severities [17, 233–235]. In our research, we conducted a Welch t-test to test if the HMPDM values in the healthy and depressive groups are significantly different; the significance levels are set at the 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 Figure 4.5), playing a key role in depression presence prediction. Additionally, the variance of $HMPDM_{17}$ increased in participants suffering from depression, while the median of the $HMPDM_{17}$ for a healthy subject was higher in healthy subjects. In depression severity classification, the $HMPDM_{17}$ of female participants was a reliable indicator when predicting depression levels, such as healthy vs. moderate, healthy vs. severe, mild vs. moderate and moderate vs. severe. For instance, Figure 4.6 shows that the healthy group has a higher median of the $HMPDM_{17}$. Further investigation is needed to fully understand the role of each audio feature in depression presence detection and level classification.

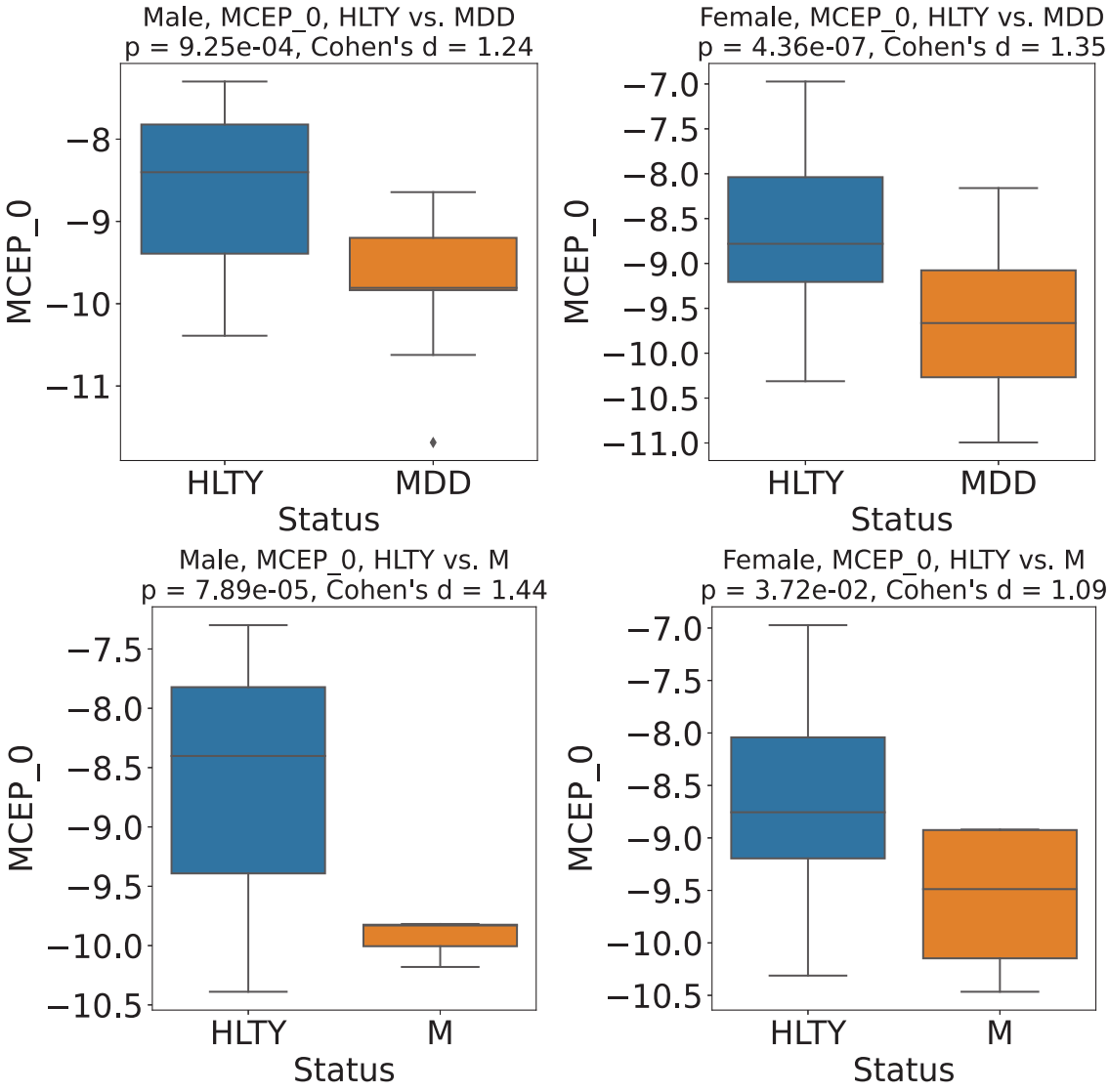

Figure 4.3: The MCEP_0 of male and female subjects in healthy vs. depressive, healthy vs. mild (HLTY: Healthy, MDD: Major Depressive Disorder, M: Mild)

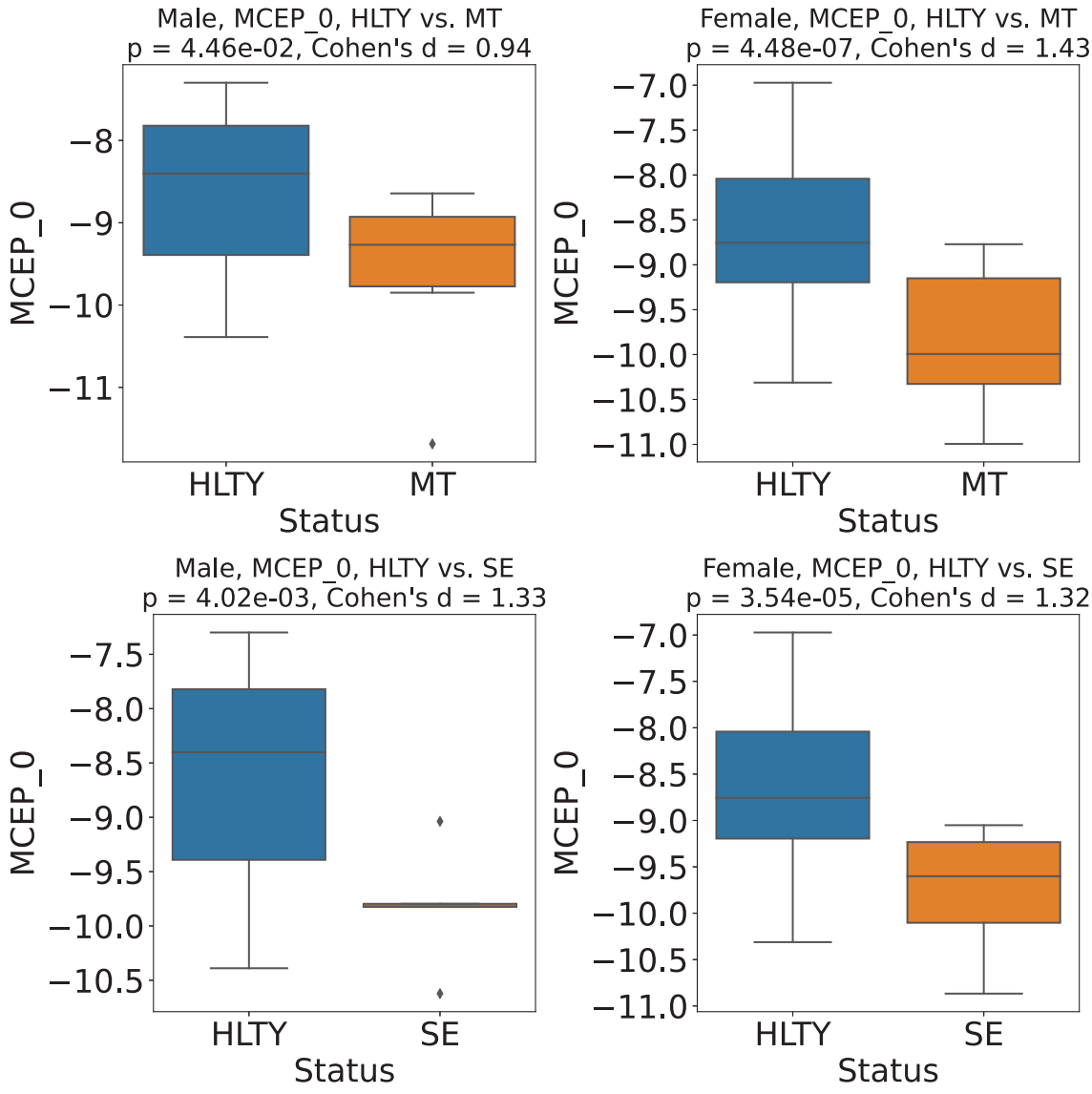

Figure 4.4: The $MCEP_0$ of male and female subjects in healthy vs. moderate, and healthy vs. severe. (HLTY: Healthy, MDD: Major Depressive Disorder, MT: Moderate, SE: Severe)

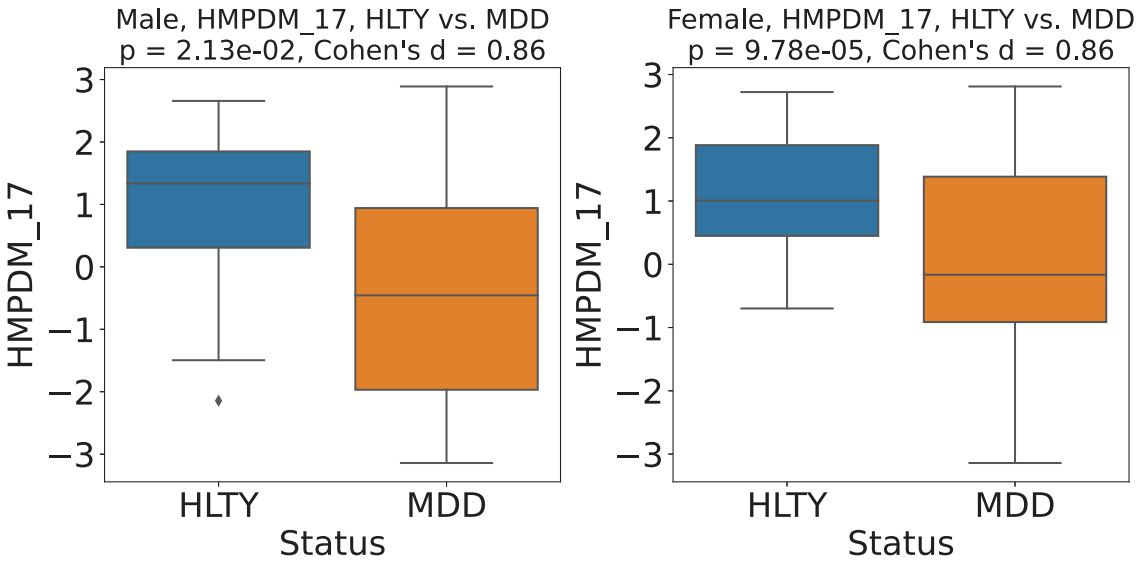

Figure 4.5: The HMPDM_{17} in healthy vs. depressive. (HLTY: Healthy, MDD: Major Depressive Disorder)

4.4.3 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 [236–239]. 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 fundamental frequency (F_0), MCEP, and HMPDM can 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. [240]. 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

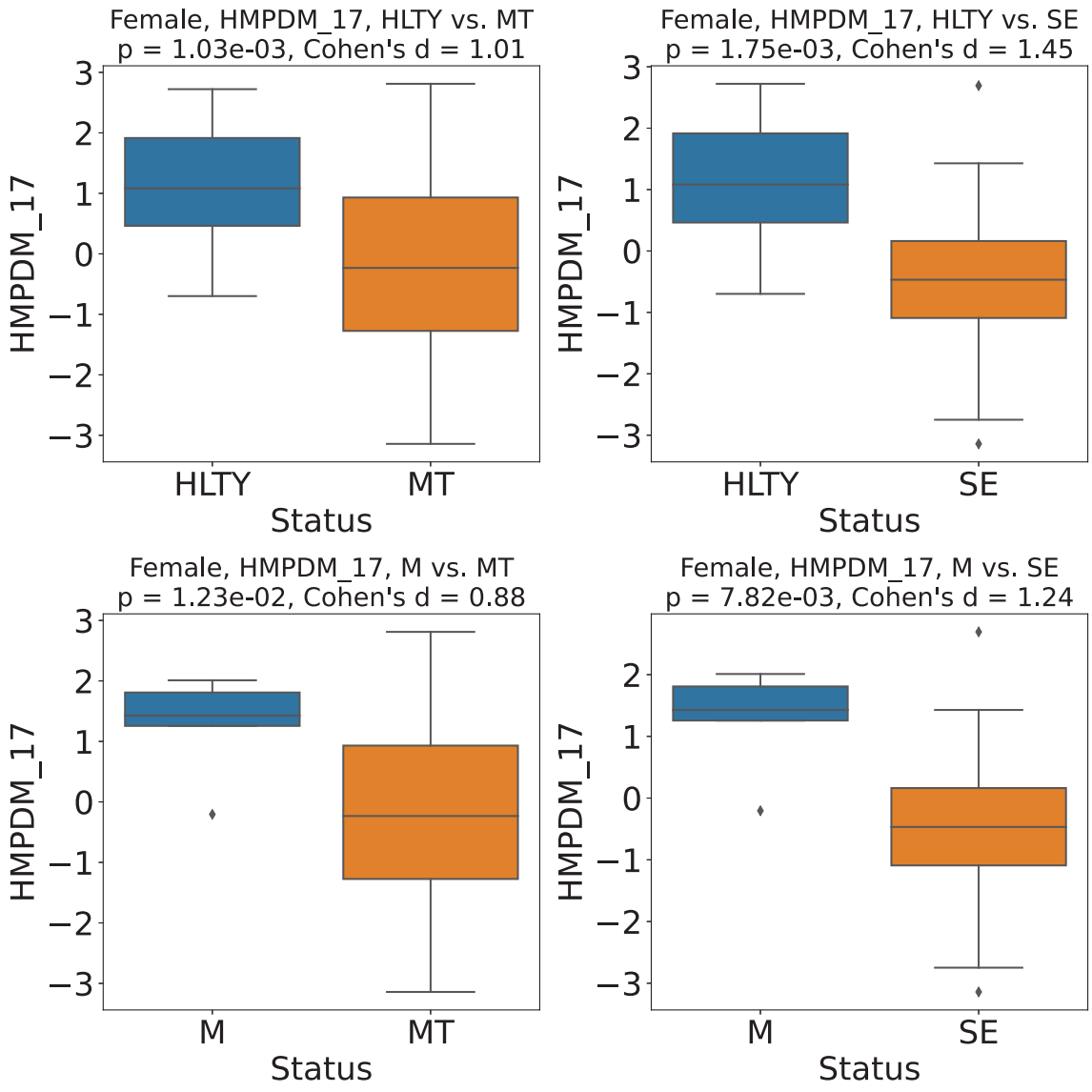

Figure 4.6: The HMPDM_{17} of the female subjects in healthy vs. moderate, healthy vs. severe, mild vs. moderate and mild vs. severe. (HLTY: Healthy, M: Mild, MT: Moderate, SE: Severe)

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 the depression probability and its explanations, which shows a series of features that increased (red) and decreased (blue) the depression risk. Based on the diagnosis of doctors, we divided the dataset into two categories; one being healthy and the other being depressive. The audio features were extracted and processed using the method described in Section 4.3.3. 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 4.4.1. 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 4.4.2, we only investigated the difference between each preprocessed audio feature (processed by the neighbourhood top-N elements method) in the healthy and depressive groups. Without a meaningful explanation, the output probability of the model may be hard to explain; by presenting the depression probability as a cumulative process, the reason for the prediction made by the model becomes clear.

The increase in the depression probability of test examples shown in Figures 4.7-4.10 is driven by audio features. The probability explanation bar in Figures 4.7-4.10 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, and a few primarily drive the probability of 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 unpredictable features that are unforeseen as effective for depression prediction. For some of these features with high contributions, it is beneficial to investigate further how they relate to depression risk. High contribution features can be used to alert psychiatrists to notice some implicit signals with depression quickly, and they are likely to be a proxy that conveys potential negative mental states or emotions.

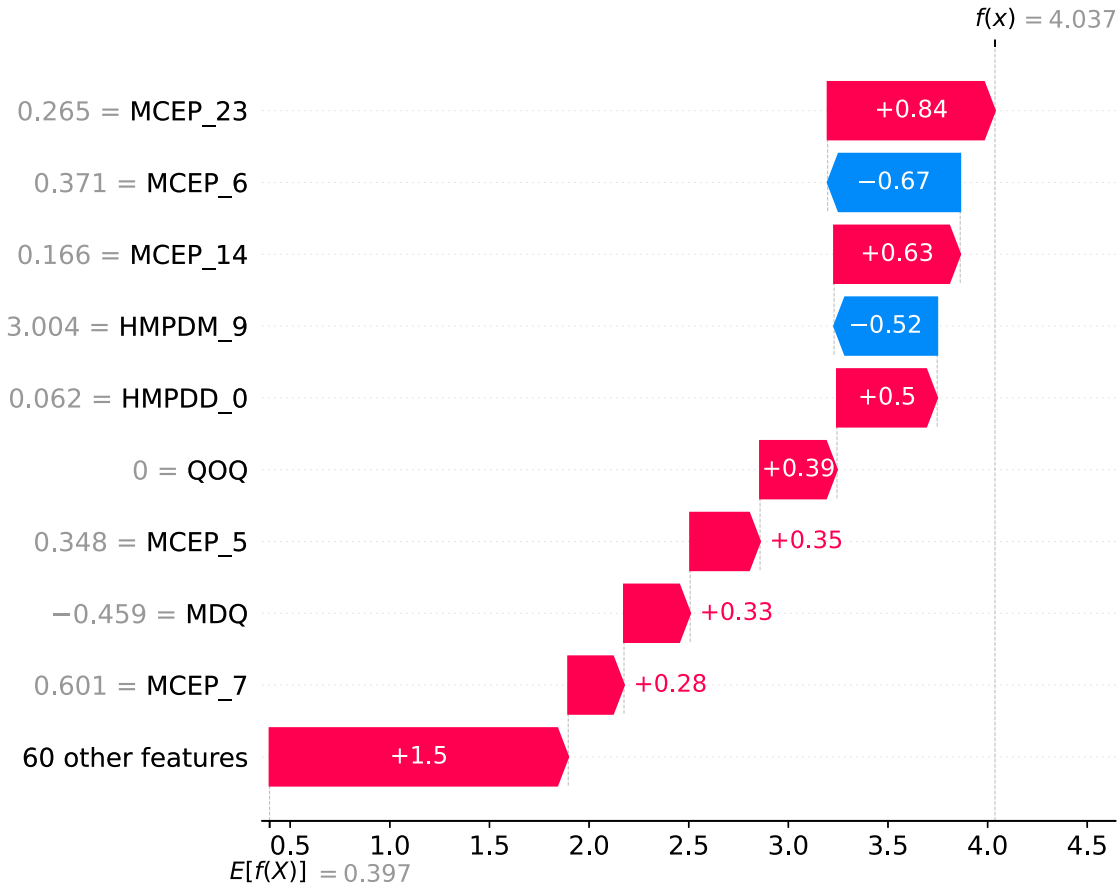

Figure 4.7: The contribution of audio features when making inferences about depressive individual S001.

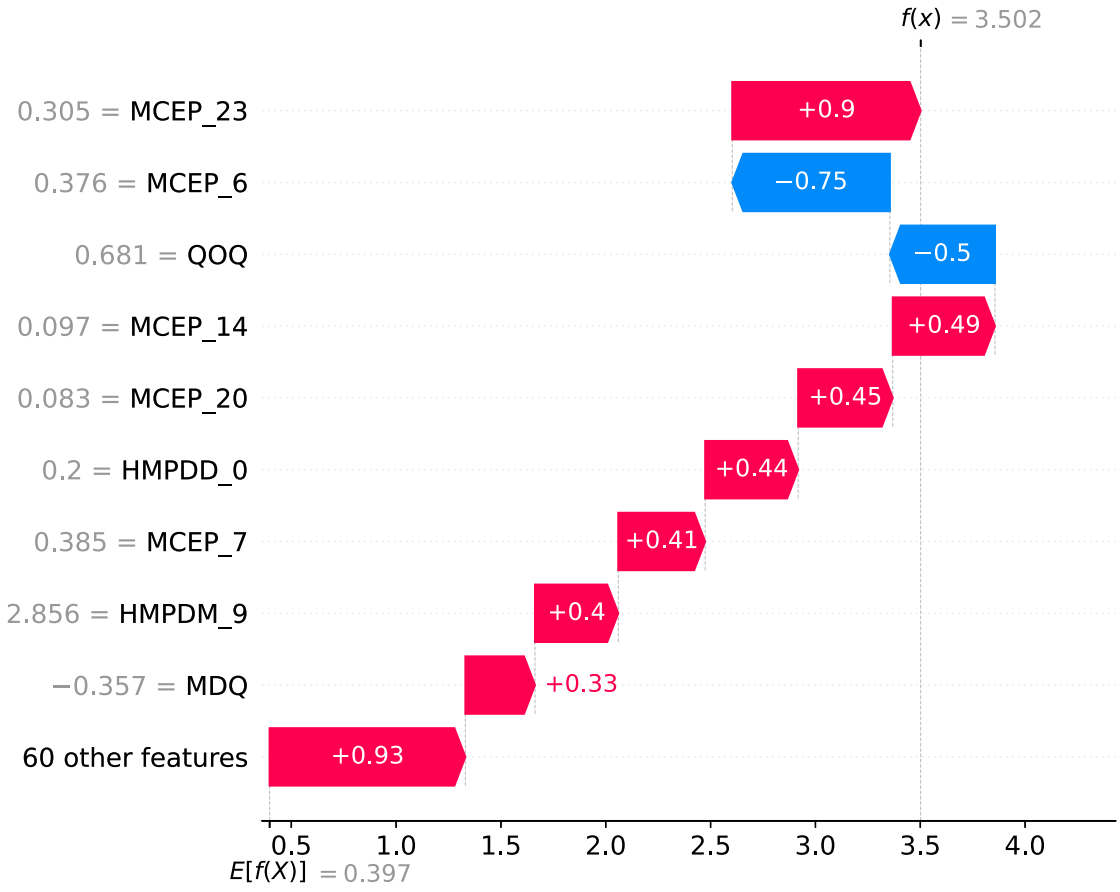

Figure 4.8: The contribution of audio features when making inferences about depressive individual S056.

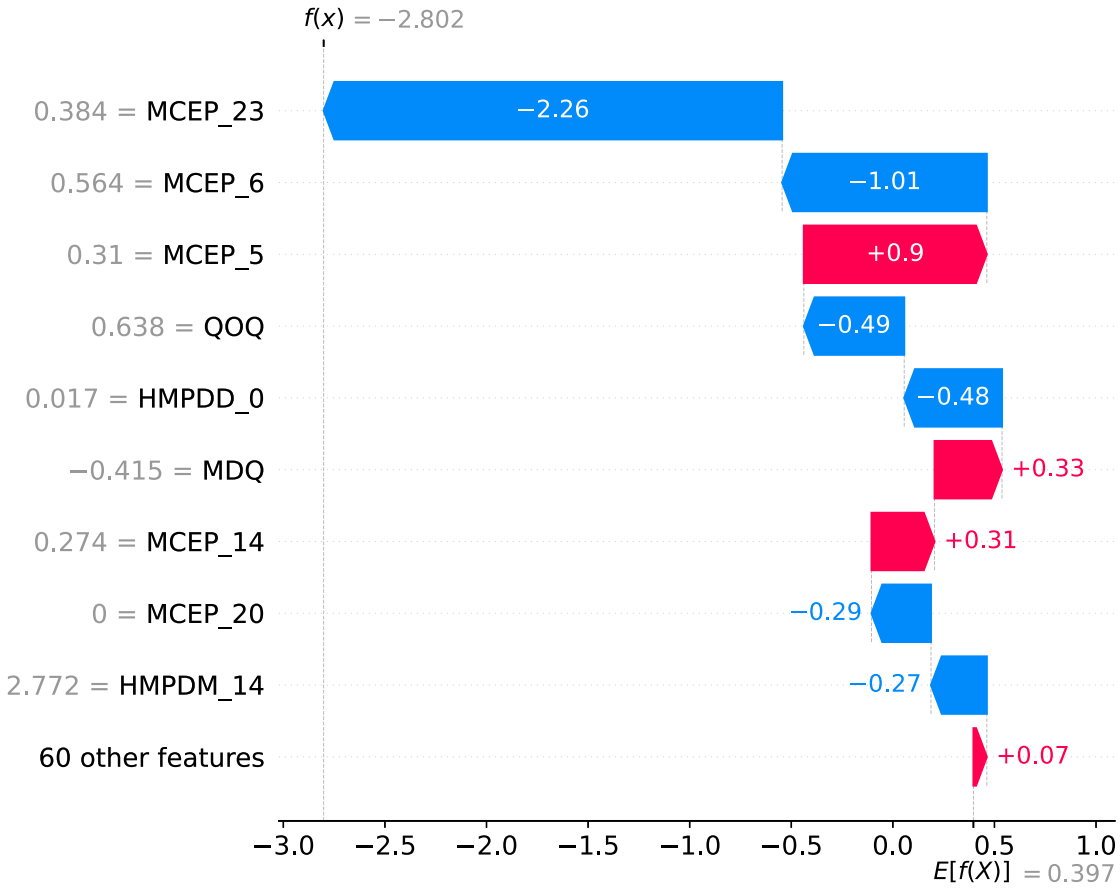

Figure 4.9: The contribution of audio features when making inferences about healthy individual S038.

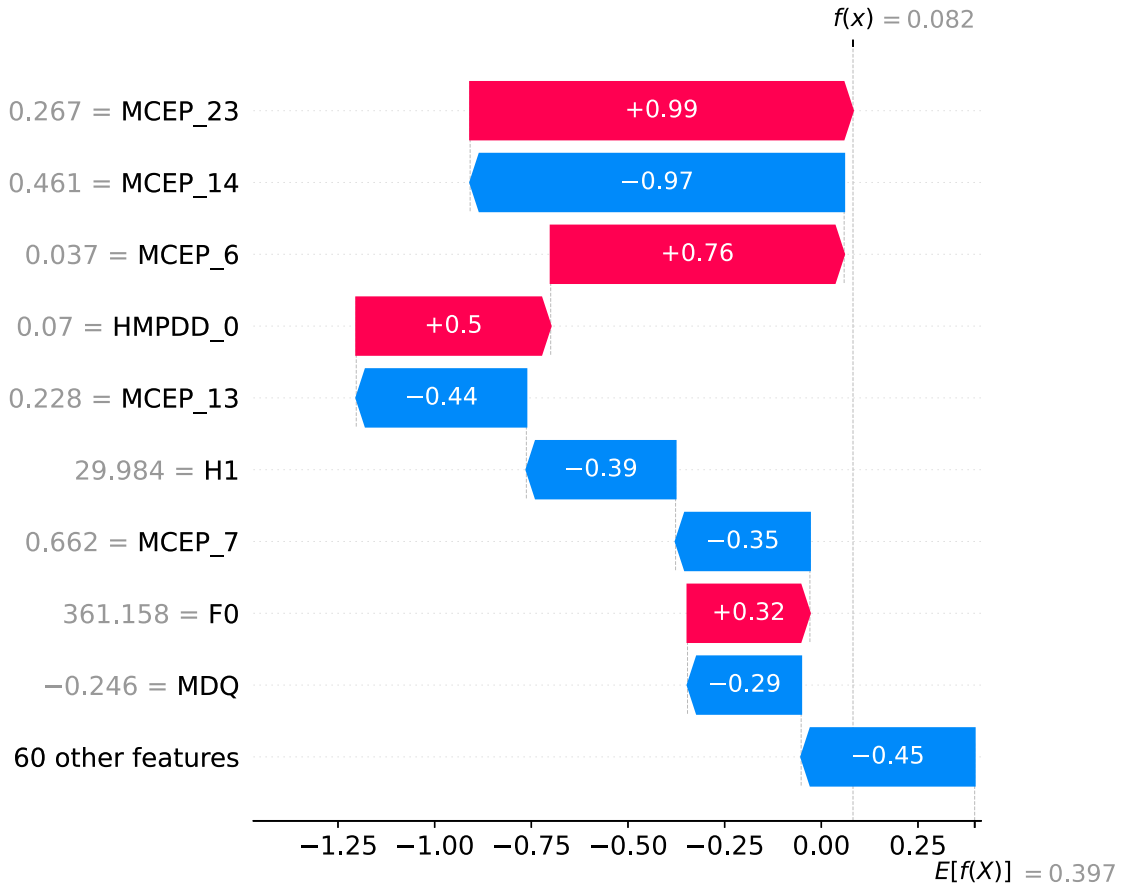

Figure 4.10: The contribution of audio features when making inferences about the healthy individual S084.

4.5 Conclusions

Open datasets are valuable for both research and the industrial community. Collecting and annotating clinical interviews with psychiatrists diagnoses is labour-intensive and requires expertise in psychology. Labelled depression interview datasets of authentic patients 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 presence and level classification on the dataset. The models achieved micro-average F_1 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 are evaluated by the nested cross-validation method and tested on the independent test set; these results demonstrate that automated depression diagnosis based on interviews in Chinese is feasible, while there is still some room for improvement. Finally, we conducted the statistical analysis in two different methods. Subjects were divided into groups based on depression severity, and intragroup feature analyses were completed. We confirmed acoustic features such as formant frequency F_0, MCEP_0, and high-order HMPDM significantly impact the ability to distinguish between depressed and non-depressed individuals.

Moreover, a novel visualization method is used to illustrate the high-impact audio features in depression detection, which further illuminated the black-box nature of our proposed models and provided a remarkable indicator for physicians to pay attention to. We anticipate the release of this dataset will motivate more researchers to focus

on 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 hope other researchers can have greater insight into AI for mental healthcare with our dataset.

Chapter 5

Conclusions, Recommendations, & Future Work

5.1 Conclusions

In Chapter 2, we reviewed 264 studies measuring acoustic, semantic, and facial landmark features to distinguish individuals with mental health disorders. Our synthesis includes significant and non-significant features across modalities, as well as guidelines for data collection and machine learning model improvement. Open-access datasets and competitions have enabled research on PTSD, bipolar disorder, and postpartum depression, but overfitting remains a concern. We encourage the collection of open datasets and code sharing for reproducibility. Conducting more research on multiple datasets may enhance model generalizability and lead to more reliable conclusions about mental health disorders. Using multimodality features for machine learning holds promise for enhancing mental health evaluations and treatment.

In Chapter 3, we presented a multimodality approach for automated depression diagnosis, using both audio and text features. Our results showed that the audio feature sequence carried information that could be used to predict depression severity, while the text features provided valuable information for depression diagnosis. Our patient-independent audio model achieved a sequence level F_1 score of 0.9870 and patient-level F_1 score of 0.9074, while the patient-independent text model achieved a

sequence level F_1 score of 0.9709 and patient-level F_1 score of 0.9245. Our experiments also revealed the best hyper-parameters for both the audio and text models. The findings provided insights for future research and assisted in model selection and hyperparameter configuration when deploying this method in clinical settings. The patient-level prediction model, obtained by a major voting algorithm, demonstrated satisfying performance. Our multimodality approach has the potential to provide an efficient and reliable tool for depression diagnosis and monitoring.

In Chapter 4, our work has demonstrated the value of open datasets in advancing research in depression diagnosis and affective computing. By collecting and analyzing a depression interview dataset labelled by clinicians, we have provided a valuable and abundant resource for other researchers working in related fields. Our baseline models achieved promising results for automated depression diagnosis based on interviews in Chinese, highlighting the feasibility of this approach. Through statistical analysis, we have confirmed the significance of specific acoustic features in depression detection and introduced a novel visualization method to aid physicians in identifying these features. We believe that the release of our dataset will inspire further research and facilitate the development of more accurate and accessible depression diagnosis tools. We hope our work contributes to a greater understanding of the potential of AI in mental healthcare and provides a benchmark for future studies in this area.

5.2 Future Work

There are still some problems should be solved, including:

For the multimodality depression detection model:

- Investigating the representation of audio/text features during the whole interview: This could help to improve the patient-level prediction by allowing the model to make predictions based on a digest of audio/text features, which could be more practical and efficient in clinical settings.

- Applying the proposed method to other mental health disorders: While the proposed multimodality approach was specifically designed for automated depression detection, it has the potential to be applied to other mental health disorders as well. It would be valuable to explore how the method performs on other disorders, such as anxiety or PTSD, and to assess whether any modifications are necessary to adapt it to these conditions. For instance, in the case of anxiety disorders, adjustments to the model's feature extraction methods may be needed to better capture anxiety-related features, such as changes in pitch, intensity, speech rate, voice quality, prosody, and silences or pauses. Similarly, for PTSD, consideration may need to be given to including a wider range of emotional and psychological states, as well as language features related to traumatic experiences. For instance, we should take lexical choice, narrative structure, emotional words, temporal references, and social interaction patterns into account. By incorporating these language features into the analysis, we can develop more comprehensive models for detecting PTSD and understanding the experiences of individuals affected by trauma. Therefore, we plan to further investigate approaches specialized to these different disorders and optimize our method based on experimental results.

For the depression interview corpus:

- Expanding the scope of labelled clinical interview datasets: While the authors have contirbuted by collecting and analyzing a depression interview dataset labelled by clinicians, there remains an insufficient quantity of such datasets available for research purposes. In the future, we aim to expand the availability of labelled depression interview datasets by collecting data from diverse populations, various demographics, cultural backgrounds, and clinical profiles. Additionally, we will aggregate data from multiple sources, including healthcare facilities, research institutions, and online platforms, to create larger and

more representative datasets for training and validating automated depression diagnosis models. By addressing these issues, we can enhance the robustness and generalizability of automated depression diagnosis models, ultimately improving effectiveness in clinical practice.

- Visualizing high-impact features: We used a novel visualization method to illustrate the high-impact audio features in depression detection. Future researchers can explore other visualization techniques to gain deeper insights into the features that impact depression diagnosis and to provide physicians with more useful indicators to pay attention to.

For the transformer-based model, our focus will remain on refining and broadening the domain-specific language model. We plan to further train the model using extensive mental health-related datasets covering a broader spectrum of topics, including autism and post-traumatic stress disorder (PTSD). This expanded training will enable the language model to be effectively utilized for the diagnosis of various mental disorders beyond its current scope.

Bibliography

[1] D. C. Mohr *et al.*, "Perceived barriers to psychological treatments and their relationship to depression," *Journal of clinical psychology*, vol. 66, no. 4, pp. 394–409, 2010, Publisher: Wiley Online Library.

[2] J. P. Docherty, "Barriers to the diagnosis of depression in primary care," *Journal of clinical psychiatry*, vol. 58, no. 1, pp. 5–10, 1997, Publisher: [Memphis, Tenn., Physicians Postgraduate Press].

[3] N. Byatt, T. A. M. Simas, R. S. Lundquist, J. V. Johnson, and D. M. Ziedonis, "Strategies for improving perinatal depression treatment in North American outpatient obstetric settings," *Journal of Psychosomatic Obstetrics & Gynecology*, vol. 33, no. 4, pp. 143–161, 2012, Publisher: Taylor & Francis.

[4] J. Han *et al.*, "Deep Learning for Mobile Mental Health: Challenges and Recent Advances," Institute of Electrical and Electronics Engineers, 2021.

[5] N. Cummins, F. Matcham, J. Klapper, and B. Schuller, "Artificial intelligence to aid the detection of mood disorders," in *Artificial Intelligence in Precision Health*, Elsevier, 2020, pp. 231–255.

[6] 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.

[7] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 668–680, 2017, Publisher: IEEE.

[8] A. R. Daros, K. K. Zakzanis, and A. Ruocco, "Facial emotion recognition in borderline personality disorder," *Psychological Medicine*, vol. 43, no. 9, pp. 1953–1963, 2013, Publisher: Cambridge University Press.

[9] Q Zhao *et al.*, "Early perceptual anomaly of negative facial expression in depression: An event-related potential study," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 45, no. 6, pp. 435–443, 2015, Publisher: Elsevier.

[10] N. Seneviratne, J. R. Williamson, A. C. Lammert, T. F. Quatieri, and C. Y. Espy-Wilson, "Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression.," in *INTERSPEECH*, 2020, pp. 4551–4555.

[11] Z. Zhao *et al.*, "Hybrid network feature extraction for depression assessment from speech," 2020.

[12] G. Kiss and K. Vicsi, "Mono-and multi-lingual depression prediction based on speech processing," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 919–935, 2017, Publisher: Springer.

[13] S. Dham, A. Sharma, and A. Dhall, "Depression scale recognition from audio, visual and text analysis," *arXiv preprint arXiv:1709.05865*, 2017.

[14] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, p. 103 107, 2022, Publisher: Elsevier.

[15] H. Jiang *et al.*, "Detecting depression using an ensemble logistic regression model based on multiple speech features," *Computational and mathematical methods in medicine*, vol. 2018, 2018, Publisher: Hindawi.

[16] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, "Audio based depression detection using Convolutional Autoencoder," *Expert Systems with Applications*, vol. 189, p. 116 076, 2022, Publisher: Elsevier.

[17] A. Pampouchidou *et al.*, "Facial geometry and speech analysis for depression detection," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2017, pp. 1433–1436.

[18] A. Maxhuni, A. Muñoz-Meléndez, V. Osmani, H. Perez, O. Mayora, and E. F. Morales, "Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients," *Pervasive and Mobile Computing*, vol. 31, pp. 50–66, 2016, Publisher: Elsevier.

[19] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[20] M. A. Rahaman, J. Chen, Z. Fu, N. Lewis, A. Iraji, and V. D. Calhoun, "Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 3267–3272.

[21] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Computers in Biology and Medicine*, vol. 135, p. 104 499, 2021, Publisher: Elsevier.

[22] S. Paul, S. K. Jandhyala, and T. Basu, "Early Detection of Signs of Anorexia and Depression Over Social Media using Effective Machine Learning Frameworks.," in *CLEF (Working notes)*, 2018.

[23] A. M. Chekroud *et al.*, "Cross-trial prediction of treatment outcome in depression: A machine learning approach," *The Lancet Psychiatry*, vol. 3, no. 3, pp. 243–250, 2016, Publisher: Elsevier.

[24] A. Mira, J. Bretón-López, A. García-Palacios, S. Quero, R. M. Baños, and C. Botella, "An Internet-based program for depressive symptoms using human and automated support: A randomized controlled trial," *Neuropsychiatric disease and treatment*, vol. 13, p. 987, 2017, Publisher: Dove Press.

[25] T. R. Insel, "The NIMH research domain criteria (RDoC) project: Precision medicine for psychiatry," *American Journal of Psychiatry*, vol. 171, no. 4, pp. 395–397, 2014, Publisher: Am Psychiatric Assoc.

[26] D. Bzdok and A. Meyer-Lindenberg, "Machine learning for precision psychiatry: Opportunities and challenges," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 3, no. 3, pp. 223–230, 2018, Publisher: Elsevier.

[27] S. Vieira, W. H. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neuroscience & Biobehavioral Reviews*, vol. 74, pp. 58–75, 2017, Publisher: Elsevier.

[28] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *NeuroImage: Clinical*, vol. 17, pp. 16–23, 2018, Publisher: Elsevier.

[29] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015, Publisher: IEEE.

[30] L. He, D. Jiang, and H. Sahli, "Multimodal depression recognition with dynamic visual and audio cues," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2015, pp. 260–266.

[31] G. Shen *et al.*, "Depression detection via harvesting social media: A multimodal dictionary learning solution.," in *IJCAI*, 2017, pp. 3838–3844.

[32] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.

[33] S. Koelstra *et al.*, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011, Publisher: IEEE.

[34] S. Koelstra, C. Mühl, and I. Patras, "EEG analysis for implicit tagging of video data," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, IEEE, 2009, pp. 1–6.

[35] Z. Zhang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3438–3446.

[36] L. Andrade *et al.*, "The epidemiology of major depressive episodes: Results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys," *International journal of methods in psychiatric research*, vol. 12, no. 1, pp. 3–21, 2003, Publisher: Wiley Online Library.

[37] J. S. Girgus, K. Yang, and C. V. Ferri, "The gender difference in depression: Are elderly women at greater risk for depression than elderly men?" *Geriatrics*, vol. 2, no. 4, p. 35, 2017, Publisher: MDPI.

[38] J. J. Schuch, A. M. Roest, W. A. Nolen, B. W. Penninx, and P. De Jonge, "Gender differences in major depressive disorder: Results from the Netherlands study of depression and anxiety," *Journal of affective disorders*, vol. 156, pp. 156–163, 2014, Publisher: Elsevier.

[39] W. Gao, S. Ping, and X. Liu, "Gender differences in depression, anxiety, and stress among college students: A longitudinal study from China," *Journal of affective disorders*, vol. 263, pp. 292–300, 2020, Publisher: Elsevier.

[40] P. R. Albert, "Why is depression more prevalent in women?" *Journal of psychiatry & neuroscience: JPN*, vol. 40, no. 4, p. 219, 2015, Publisher: Canadian Medical Association.

[41] C. J. Murray and A. D. Lopez, "Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study," *The lancet*, vol. 349, no. 9063, pp. 1436–1442, 1997, Publisher: Elsevier.

[42] J. Olesen *et al.*, "The economic cost of brain disorders in Europe," *European journal of neurology*, vol. 19, no. 1, pp. 155–162, 2012, Publisher: Wiley Online Library.

[43] W. F. Stewart, J. A. Ricci, E. Chee, S. R. Hahn, and D. Morganstein, "Cost of lost productive work time among US workers with depression," *Jama*, vol. 289, no. 23, pp. 3135–3144, 2003, Publisher: American Medical Association.

[44] W. H. Organization and others, "Preventing suicide: A global imperative," 2014, Publisher: World Health Organization.

[45] K. Hawton, C. C. i Comabella, C. Haw, and K. Saunders, "Risk factors for suicide in individuals with depression: A systematic review," *Journal of affective disorders*, vol. 147, no. 1-3, pp. 17–28, 2013, Publisher: Elsevier.

[46] J.-P. Lépine and M. Briley, "The increasing burden of depression," *Neuropsychiatric disease and treatment*, vol. 7, no. Suppl 1, p. 3, 2011, Publisher: Dove Press.

[47] T. E. Joiner Jr, J. S. Brown, and L. R. Wingate, "The psychology and neurobiology of suicidal behavior," *Annu. Rev. Psychol.*, vol. 56, pp. 287–314, 2005, Publisher: Annual Reviews.

[48] A. McGirr *et al.*, "An examination of DSM-IV depressive symptoms and risk for suicide completion in major depressive disorder: A psychological autopsy study," *Journal of Affective Disorders*, vol. 97, no. 1-3, pp. 203–209, 2007, Publisher: Elsevier.

[49] S. Evans-Lacko *et al.*, "Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the WHO World Mental Health (WMH) surveys," *Psychological medicine*, vol. 48, no. 9, pp. 1560–1571, 2018, Publisher: Cambridge University Press.

[50] R. Freedman *et al.*, *The initial field trials of DSM-5: New blooms and old thorns*, Issue: 1 Pages: 1–5 Publication Title: American Journal of Psychiatry Volume: 170, 2013.

[51] D. A. Regier *et al.*, "DSM-5 field trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses," *American journal of psychiatry*, vol. 170, no. 1, pp. 59–70, 2013, Publisher: Am Psychiatric Assoc.

[52] C. C. Joyal, J.-L. Dubreucq, C. Gendron, and F. Millaud, "Major mental disorders and violence: A critical update," *Current psychiatry reviews*, vol. 3, no. 1, pp. 33–50, 2007, Publisher: Bentham Science Publishers.

[53] S. Montgomery and M. Åsberg, *A new depression scale designed to be sensitive to change*. Acad. Department of Psychiatry, Guy's Hospital, 1977.

[54] 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. DOI: 10.1109/ICASSP.2014.6853739.

[55] E. Kraepelin, "Manic depressive insanity and paranoia," *The Journal of Nervous and Mental Disease*, vol. 53, no. 4, p. 350, 1921, Publisher: LWW.

[56] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020, Publisher: Wiley Online Library.

[57] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Communication*, vol. 75, pp. 27–49, 2015, Publisher: Elsevier.

[58] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Probabilistic acoustic volume analysis for speech affected by depression," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[59] S. Harati, A. Crowell, H. Mayberg, and S. Nemati, "Depression severity classification from speech emotion," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2018, pp. 5763–5766.

[60] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller, "Enhancing speech-based depression detection through gender dependent vowel-level formant features," in *Conference on artificial intelligence in medicine in Europe*, Springer, 2017, pp. 209–214.

[61] M. R. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," in *2016 IEEE spoken language technology workshop (SLT)*, IEEE, 2016, pp. 136–143.

[62] K. Vicsi, D. Sztahó, and G. Kiss, "Examination of the sensitivity of acoustic-phonetic parameters of speech to depression," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, IEEE, 2012, pp. 511–515.

[63] G. Kiss and K. Vicsi, "Comparison of read and spontaneous speech in case of automatic detection of depression," in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, IEEE, 2017, pp. 000 213–000 218.

[64] B. Stasak, Z. Huang, D. Joachim, and J. Epps, "Automatic elicitation compliance for short-duration speech based depression detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7283–7287.

[65] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and PTSD.," in *Interspeech*, 2013, pp. 847–851.

[66] S. Marmor, K. J. Horvath, K. O. Lim, and S. Misono, "Voice problems and depression among adults in the U nited S tates," *The Laryngoscope*, vol. 126, no. 8, pp. 1859–1864, 2016, Publisher: Wiley Online Library.

[67] M. Kumar, M. Dredze, G. Coppersmith, and M. De Choudhury, "Detecting changes in suicide content manifested in social media following celebrity suicides," in *Proceedings of the 26th ACM conference on Hypertext & Social Media*, 2015, pp. 85–94.

[68] I. Pirina and Ç. Çöltekin, "Identifying depression on reddit: The effect of training data," in *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*, 2018, pp. 9–12.

[69] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," *arXiv preprint arXiv:1709.01848*, 2017.

[70] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes, "Detecting depression in social media using fine-grained emotions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 1481–1486.

[71] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: An integrative review," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017, Publisher: Elsevier.

[72] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in reddit social media forum," *IEEE Access*, vol. 7, pp. 44 883–44 893, 2019, Publisher: IEEE.

[73] M. De Choudhury and S. De, "Mental health discourse on reddit: Self-disclosure, social support, and anonymity," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[74] N. S. Alghamdi, H. A. H. Mahmoud, A. Abraham, S. A. Alanazi, and L. García-Hernández, "Predicting depression symptoms in an Arabic psychological forum," *IEEE Access*, vol. 8, pp. 57 317–57 334, 2020, Publisher: IEEE.

[75] A. H. Yazdavar *et al.*, "Semi-supervised approach to monitoring clinical depressive symptoms in social media," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 1191–1198.

[76] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "Depressionnet: A novel summarization boosted deep framework for depression detection on social media," *arXiv preprint arXiv:2105.10878*, 2021.

[77] K. Yang, T. Zhang, and S. Ananiadou, "A mental state knowledge–aware and contrastive network for early stress and depression detection on social media," *Information Processing & Management*, vol. 59, no. 4, p. 102 961, 2022.

[78] J. S. Lara, M. E. Aragón, F. A. González, and M. Montes-y Gómez, "Deep bag-of-sub-emotions for depression detection in social media," in *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*, Springer International Publishing, 2021, pp. 60–72.

[79] J. S. L. Figuerêdo, A. L. L. Maia, and R. T. Calumby, "Early depression detection in social media based on deep learning and underlying emotions," *Online Social Networks and Media*, vol. 31, p. 100 225, 2022.

[80] M. Stankevich, V. Isakov, D. Devyatkin, and I. V. Smirnov, "Feature engineering for depression detection in social media.," in *ICPRAM*, 2018, pp. 426–431.

[81] J. de Jesús Titla-Tlatelpa, R. M. Ortega-Mendoza, M. Montes-y Gómez, and L. Villaseñor-Pineda, "A profile-based sentiment-aware approach for depression detection in social media," *EPJ data science*, vol. 10, no. 1, p. 54, 2021.

[82] Z. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, "Mha: A multimodal hierarchical attention model for depression detection in social media," *Health Information Science and Systems*, vol. 11, no. 1, p. 6, 2023.

[83] J. Cha, S. Kim, and E. Park, "A lexicon-based approach to examine depression detection in social media: The case of twitter and university community," *Humanities and Social Sciences Communications*, vol. 9, no. 1, pp. 1–10, 2022.

[84] B. Cui, J. Wang, H. Lin, Y. Zhang, L. Yang, and B. Xu, "Emotion-based reinforcement attention network for depression detection on social media: Algorithm development and validation," *JMIR Medical Informatics*, vol. 10, no. 8, e37818, 2022.

[85] Z. Guo, N. Ding, M. Zhai, Z. Zhang, and Z. Li, "Leveraging domain knowledge to improve depression detection on chinese social media," *IEEE Transactions on Computational Social Systems*, 2023.

[86] S. H. Hosseini-Saravani, S. Besharati, H. Calvo, and A. Gelbukh, "Depression detection in social media using a psychoanalytical technique for feature extraction and a cognitive based classifier," in *Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Mexico City, Mexico, October 12–17, 2020, Proceedings, Part II*, Springer, 2020, pp. 282–292.

[87] F. Ramiandrisoa and J. Mothe, "Early detection of depression and anorexia from social media: A machine learning approach," in *Circle 2020*, vol. 2621, 2020.

[88] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol. 25, no. 1, pp. 281–304, 2022.

[89] A. Hartanto, F. Y. Quek, G. Y. Tng, and J. C. Yong, "Does social media use increase depressive symptoms? a reverse causation perspective," *Frontiers in Psychiatry*, vol. 12, p. 641 934, 2021.

[90] B. A. Primack *et al.*, "Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among us young adults," *Computers in human behavior*, vol. 69, pp. 1–9, 2017.

[91] B. A. Primack, A. Shensa, J. E. Sidani, C. G. Escobar-Viera, and M. J. Fine, "Temporal associations between social media use and depression," *American journal of preventive medicine*, vol. 60, no. 2, pp. 179–188, 2021.

[92] K. Puukko, L. Hietajärvi, E. Maksniemi, K. Alho, and K. Salmela-Aro, "Social media use and depressive symptoms—a longitudinal study from early to late adolescence," *International journal of environmental research and public health*, vol. 17, no. 16, p. 5921, 2020.

[93] I. E. Thorisdottir, R. Sigurvinsdottir, B. B. Asgeirsdottir, J. P. Allegrante, and I. D. Sigfusdottir, "Active and passive social media use and symptoms of anxiety and depressed mood among icelandic adolescents," *Cyberpsychology, Behavior, and Social Networking*, vol. 22, no. 8, pp. 535–542, 2019.

[94] S. Cunningham, C. C. Hudson, and K. Harkness, "Social media and depression symptoms: A meta-analysis," *Research on child and adolescent psychopathology*, vol. 49, no. 2, pp. 241–253, 2021, Publisher: Springer.

[95] A. Shensa, C. G. Escobar-Viera, J. E. Sidani, N. D. Bowman, M. P. Marshal, and B. A. Primack, "Problematic social media use and depressive symptoms among US young adults: A nationally-representative study," *Social science & medicine*, vol. 182, pp. 150–157, 2017, Publisher: Elsevier.

[96] H. C. Woods and H. Scott, "# Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem," *Journal of adolescence*, vol. 51, pp. 41–49, 2016, Publisher: Elsevier.

[97] E. J. Ivie, A. Pettitt, L. J. Moses, and N. B. Allen, "A meta-analysis of the association between adolescent social media use and depressive symptoms," *Journal of affective disorders*, vol. 275, pp. 165–174, 2020, Publisher: Elsevier.

[98] L. Raudsepp and K. Kais, "Longitudinal associations between problematic social media use and depressive symptoms in adolescent girls," *Preventive medicine reports*, vol. 15, p. 100 925, 2019, Publisher: Elsevier.

[99] B. Zhong, Y. Huang, and Q. Liu, "Mental health toll from the coronavirus: Social media usage reveals Wuhan residents' depression and secondary trauma in the COVID-19 outbreak," *Computers in human behavior*, vol. 114, p. 106 524, 2021, Publisher: Elsevier.

[100] R. Haand and Z. Shuwang, "The relationship between social media addiction and depression: A quantitative study among university students in Khost, Afghanistan," *International Journal of Adolescence and Youth*, vol. 25, no. 1, pp. 780–786, 2020, Publisher: Taylor & Francis.

[101] J. Brailovskaia and J. Margraf, "Relationship between depression symptoms, physical activity, and addictive social media use," *Cyberpsychology, Behavior, and Social Networking*, vol. 23, no. 12, pp. 818–822, 2020, Publisher: Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New …

[102] A. Jeri-Yabar *et al.*, "Association between social media use (Twitter, Instagram, Facebook) and depressive symptoms: Are Twitter users at higher risk?" *International Journal of Social Psychiatry*, vol. 65, no. 1, pp. 14–19, 2019, Publisher: SAGE Publications Sage UK: London, England.

[103] K. Kircaburun, "Self-Esteem, Daily Internet Use and Social Media Addiction as Predictors of Depression among Turkish Adolescents.," *Journal of Education and Practice*, vol. 7, no. 24, pp. 64–72, 2016, Publisher: ERIC.

[104] J. Hussain *et al.*, "Exploring the dominant features of social media for depression detection," *Journal of Information Science*, vol. 46, no. 6, pp. 739–759, 2020.

[105] A. S. Liaw and H. N. Chua, "Depression detection on social media with user network and engagement features using machine learning methods," in *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, IEEE, 2022, pp. 1–6.

[106] M. E. Aragon, A. P. Lopez-Monroy, L.-C. G. Gonzalez-Gurrola, and M. Montes, "Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression," *IEEE Transactions on Affective Computing*, 2021.

[107] N. Vedula and S. Parthasarathy, "Emotional and linguistic cues of depression from social media," in *Proceedings of the 2017 International Conference on Digital Health*, 2017, pp. 127–136.

[108] J. Nesi *et al.*, "Emotional responses to social media experiences among adolescents: Longitudinal associations with depressive symptoms," *Journal of Clinical Child & Adolescent Psychology*, pp. 1–16, 2021.

[109] S. Ghosh and T. Anwar, "Depression intensity estimation via social media: A deep learning approach," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 6, pp. 1465–1474, 2021.

[110] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Seventh international AAAI conference on weblogs and social media*, 2013.

[111] A. Radovic, T. Gmelin, B. D. Stein, and E. Miller, "Depressed adolescents' positive and negative use of social media," *Journal of adolescence*, vol. 55, pp. 5–15, 2017, Publisher: Elsevier.

[112] A. Robinson *et al.*, "Social comparisons, social media addiction, and social interaction: An examination of specific social media behaviors related to major depressive disorder in a millennial population," *Journal of Applied Biobehavioral Research*, vol. 24, no. 1, e12158, 2019.

[113] R. Salas-Zárate, G. Alor-Hernández, M. d. P. Salas-Zárate, M. A. Paredes-Valverde, M. Bustos-López, and J. L. Sánchez-Cervantes, "Detecting depression signs on social media: A systematic literature review," in *Healthcare*, Issue: 2, vol. 10, MDPI, 2022, p. 291.

[114] D. Liu, X. L. Feng, F. Ahmed, M. Shahid, J. Guo, and others, "Detecting and measuring depression on social media using a machine learning approach: Systematic review," *JMIR Mental Health*, vol. 9, no. 3, e27244, 2022, Publisher: JMIR Publications Inc., Toronto, Canada.

[115] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes, "Detecting Depression in Social Media using Fine-Grained Emotions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 1481–1486.

[116] N. McCrae, S. Gettings, and E. Purssell, "Social media and depressive symptoms in childhood and adolescence: A systematic review," *Adolescent Research Review*, vol. 2, no. 4, pp. 315–330, 2017, Publisher: Springer.

[117] T. Heffer, M. Good, O. Daly, E. MacDonell, and T. Willoughby, "The longitudinal association between social-media use and depressive symptoms among adolescents and young adults: An empirical reply to Twenge et al.(2018)," *Clinical Psychological Science*, vol. 7, no. 3, pp. 462–470, 2019, Publisher: Sage Publications Sage CA: Los Angeles, CA.

[118] Z. Peng, Q. Hu, and J. Dang, "Multi-kernel SVM based depression recognition using social media data," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 43–57, 2019, Publisher: Springer.

[119] M. M. Aldarwish and H. F. Ahmad, "Predicting depression levels using social media posts," in *2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS)*, IEEE, 2017, pp. 277–280.

[120] S Smys and J. S. Raj, "Analysis of deep learning techniques for early detection of depression on social media network-a comparative study," *Journal of trends in Computer Science and Smart technology (TCSST)*, vol. 3, no. 01, pp. 24–39, 2021.

[121] A.-M. Bucur and L. P. Dinu, "Detecting early onset of depression from social media text using learned confidence scores," *arXiv preprint arXiv:2011.01695*, 2020.

[122] S Kayalvizhi and D Thenmozhi, "Data set creation and empirical analysis for detecting signs of depression from social media postings," *arXiv preprint arXiv:2202.03047*, 2022.

[123] P. Mann, A. Paes, and E. H. Matsushima, "See and read: Detecting depression symptoms in higher education students using multimodal social media data," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 440–451.

[124] F. Sadeque, D. Xu, and S. Bethard, "Measuring the latency of depression detection in social media," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 495–503.

[125] I. Fatima, B. U. D. Abbasi, S. Khan, M. Al-Saeed, H. F. Ahmad, and R. Mumtaz, "Prediction of postpartum depression using machine learning techniques from social media text," *Expert Systems*, vol. 36, no. 4, e12409, 2019, Publisher: Wiley Online Library.

[126] K. Katchapakirin, K. Wongpatikaseree, P. Yomaboot, and Y. Kaewpitakkun, "Facebook social media for depression detection in the Thai community," in *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, IEEE, 2018, pp. 1–6.

[127] Q. Wang, H. Yang, and Y. Yu, "Facial expression video analysis for depression detection in Chinese patients," *Journal of Visual Communication and Image Representation*, vol. 57, pp. 228–233, 2018, Publisher: Elsevier.

[128] Y. Hao, Y. Cao, B. Li, and M. Rahman, "Depression recognition based on text and facial expression," in *International Symposium on Artificial Intelligence and Robotics 2021*, vol. 11884, SPIE, 2021, pp. 513–522.

[129] D. Li, H. Chaudhary, and Z. Zhang, "Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining," *International Journal of Environmental Research and Public Health*, vol. 17, no. 14, p. 4988, 2020, Publisher: MDPI.

[130] Z. Liu, X. Yuan, Y. Li, Z. Shangguan, L. Zhou, and B. Hu, "Pra-net: Part-and-relation attention network for depression recognition from facial expression," *Computers in Biology and Medicine*, vol. 157, p. 106 589, 2023.

[131] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 43–50.

[132] L. Hunter, L. Roland, and A. Ferozpuri, "Emotional expression processing and depressive symptomatology: Eye-tracking reveals differential importance of lower and middle facial areas of interest," *Depression research and treatment*, vol. 2020, 2020, Publisher: Hindawi.

[133] D. S. B. A. Hamid, S. Goyal, and P. Bedi, "Integration of deep learning for improved diagnosis of depression using eeg and facial features," *Materials Today: Proceedings*, vol. 80, pp. 1965–1969, 2023.

[134] Z. Shangguan, Z. Liu, G. Li, Q. Chen, Z. Ding, and B. Hu, "Dual-stream multiple instance learning for depression detection with facial expression videos," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.

[135] Z. Dai, Q. Li, Y. Shang, and X. Wang, "Depression detection based on facial expression, audio and gait," in *2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, IEEE, vol. 6, 2023, pp. 1568–1573.

[136] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 73–80.

[137] B. Sumali, Y. Mitsukura, Y. Tazawa, and T. Kishimoto, "Facial landmark activity features for depression screening," in *2019 58th annual conference of the society of instrument and control engineers of Japan (SICE)*, IEEE, 2019, pp. 1376–1381.

[138] B. G. Dadiz and C. R. Ruiz, "Detecting depression in videos using uniformed local binary pattern on facial features," in *Computational Science and Technology: 5th ICCST 2018, Kota Kinabalu, Malaysia, 29-30 August 2018*, Springer, 2019, pp. 413–422.

[139] S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 65–71.

[140] L. Zhang, J. Driscol, X. Chen, and R. Hosseini Ghomi, "Evaluating acoustic and linguistic features of detecting depression sub-challenge dataset," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 47–53.

[141] M. Asgari, I. Shafran, and L. B. Sheeber, "Inferring clinical depression from speech and spoken utterances," in *2014 IEEE international workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, 2014, pp. 1–5.

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

[143] W. Pan *et al.*, "Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders," *PloS one*, vol. 14, no. 6, e0218172, 2019.

[144] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews," UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES, Tech. Rep., 2014.

[145] M. Valstar *et al.*, "Avec 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 3–10.

[146] M. Valstar *et al.*, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th international workshop on audio/visual emotion challenge*, 2014, pp. 3–10.

[147] M. De Hert, J. Detraux, R. Van Winkel, W. Yu, and C. U. Correll, "Metabolic and cardiovascular adverse effects associated with antipsychotic drugs," *Nature Reviews Endocrinology*, vol. 8, no. 2, pp. 114–126, 2012, Publisher: Nature Publishing Group.

[148] J. Kane, M. Aylett, I. Yanushevskaya, and C. Gobl, "Phonetic feature extraction for context-sensitive glottal source processing," *Speech Communication*, vol. 59, pp. 10–21, 2014, Publisher: Elsevier.

[149] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and PTSD.," in *Interspeech*, 2013, pp. 847–851.

[150] S. Alghowinem *et al.*, "From joyous to clinically depressed: Mood detection using spontaneous speech.," in *FLAIRS Conference*, vol. 19, 2012.

[151] D. DeVault *et al.*, "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 1061–1068.

[152] A. Hartholt *et al.*, "All together now," in *International Workshop on Intelligent Virtual Agents*, Springer, 2013, pp. 368–381.

[153] C. Burton *et al.*, "Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression," *Journal of telemedicine and telecare*, vol. 22, no. 6, pp. 348–355, 2016, Publisher: SAGE Publications Sage UK: London, England.

[154] V. Nemes, D. Nikolic, A. Barney, and P. Garrard, "A feasibility study of speech recording using a contact microphone in patients with possible or probable Alzheimer's disease to detect and quantify repetitions in a natural setting," *Alzheimer's & Dementia*, vol. 8, no. 4, P490–P491, 2012, Publisher: No longer published by Elsevier.

[155] Z. Huang, J. Epps, and D. Joachim, "Speech landmark bigrams for depression detection from naturalistic smartphone speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 5856–5860.

[156] Z. Huang, J. Epps, D. Joachim, and M. Chen, "Depression Detection from Short Utterances via Diverse Smartphones in Natural Environmental Conditions.," in *INTERSPEECH*, 2018, pp. 3393–3397.

[157] E Szabadi, C. Bradshaw, and J. Besson, "Elongation of pause-time in speech: A simple, objective measure of motor retardation in depression," *The British Journal of Psychiatry*, vol. 129, no. 6, pp. 592–597, 1976, Publisher: Cambridge University Press.

[158] H. Pérez Espinosa, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, and V. Reyez-Meza, "Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition: INAOE-BUAP's Participation at AVEC'14 Challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 49–55.

[159] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2011, pp. 2376–2379.

[160] M. Semkovska, M. Noone, M. Carton, and D. M. McLoughlin, "Measuring consistency of autobiographical memory recall in depression," *Psychiatry research*, vol. 197, no. 1-2, pp. 41–48, 2012, Publisher: Elsevier.

[161] S. Saeb, E. G. Lattie, K. P. Kording, D. C. Mohr, and others, "Mobile phone detection of semantic location and its relationship to depression and anxiety," *JMIR mHealth and uHealth*, vol. 5, no. 8, e7297, 2017, Publisher: JMIR Publications Inc., Toronto, Canada.

[162] N. Aloshban, A. Esposito, and A. Vinciarelli, "What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech," *Cognitive Computation*, vol. 14, no. 5, pp. 1585–1598, 2022, Publisher: Springer.

[163] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, e0118432, 2015, Publisher: Public Library of Science San Francisco, CA USA.

[164] I Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989, Publisher: JSTOR.

[165] F. Wang, R. Kaushal, and D. Khullar, *Should health care demand interpretable artificial intelligence or accept "black box" medicine?* Issue: 1 Pages: 59–60 Publication Title: Annals of internal medicine Volume: 172, 2020.

[166] T. P. Quinn, S. Jacobs, M. Senadeera, V. Le, and S. Coghlan, "The three ghosts of medical AI: Can the black-box present deliver?" *Artificial intelligence in medicine*, vol. 124, p. 102 158, 2022, Publisher: Elsevier.

[167] M. Sendak *et al.*, "" The human body is a black box" supporting clinical decision-making with deep learning," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 99–109.

[168] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, 2018, Publisher: ACM New York, NY, USA.

[169] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[170] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st international conference on neural information processing systems*, 2017, pp. 4768–4777.

[171] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[172] A. L. Beam, A. K. Manrai, and M. Ghassemi, "Challenges to the reproducibility of machine learning models in health care," *Jama*, vol. 323, no. 4, pp. 305–306, 2020, Publisher: American Medical Association.

[173] M. B. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi, "Reproducibility in machine learning for health research: Still a ways to go," *Science Translational Medicine*, vol. 13, no. 586, eabb1655, 2021, Publisher: American Association for the Advancement of Science.

[174] B. Custers, "Click here to consent forever: Expiry dates for informed consent," *Big Data & Society*, vol. 3, no. 1, p. 2 053 951 715 624 935, 2016, Publisher: SAGE Publications Sage UK: London, England.

[175] A. Rahman, A. Malik, S. Sikander, C. Roberts, and F. Creed, "Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: A cluster-randomised controlled trial," *The Lancet*, vol. 372, no. 9642, pp. 902–909, 2008, Publisher: Elsevier.

[176] M. G. Mazza *et al.*, "Anxiety and depression in COVID-19 survivors: Role of inflammatory and clinical predictors," *Brain, behavior, and immunity*, vol. 89, pp. 594–600, 2020, Publisher: Elsevier.

[177] N. Salari *et al.*, "Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: A systematic review and meta-analysis," *Globalization and health*, vol. 16, no. 1, pp. 1–11, 2020, Publisher: BioMed Central.

[178] R. Barzilay *et al.*, "Resilience, COVID-19-related stress, anxiety and depression during the pandemic in a large population enriched for healthcare providers," *Translational psychiatry*, vol. 10, no. 1, pp. 1–8, 2020, Publisher: Nature Publishing Group.

[179] H. C. Nguyen *et al.*, "People with suspected COVID-19 symptoms were more likely depressed and had lower health-related quality of life: The potential benefit of health literacy," *Journal of clinical medicine*, vol. 9, no. 4, p. 965, 2020, Publisher: Multidisciplinary Digital Publishing Institute.

[180] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS medicine*, vol. 3, no. 11, e442, 2006, Publisher: Public Library of Science.

[181] S. Rodrigues *et al.*, "Impact of stigma on veteran treatment seeking for depression," *American Journal of Psychiatric Rehabilitation*, vol. 17, no. 2, pp. 128–146, 2014, Publisher: Taylor & Francis.

[182] M. B. First and M. Gibbon, "The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) and the Structured Clinical Interview for DSM-IV Axis II Disorders (SCID-II).," 2004, Publisher: John Wiley & Sons Inc.

[183] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: Validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001, Publisher: Wiley Online Library.

[184] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An inventory for measuring depression," *Archives of general psychiatry*, vol. 4, no. 6, pp. 561–571, 1961, Publisher: American Medical Association.

[185] S. A. Montgomery and M. Åsberg, "A new depression scale designed to be sensitive to change," *The British journal of psychiatry*, vol. 134, no. 4, pp. 382–389, 1979, Publisher: Cambridge University Press.

[186] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004, Publisher: Taylor & Francis.

[187] R. M. Bagby, A. G. Ryder, D. R. Schuller, and M. B. Marshall, "The Hamilton Depression Rating Scale: Has the gold standard become a lead weight?" *American Journal of Psychiatry*, vol. 161, no. 12, pp. 2163–2177, 2004, Publisher: Am Psychiatric Assoc.

[188] H. C. Kraemer, D. J. Kupfer, D. E. Clarke, W. E. Narrow, and D. A. Regier, "DSM-5: How reliable is reliable enough?" *American Journal of Psychiatry*, vol. 169, no. 1, pp. 13–15, 2012, Publisher: Am Psychiatric Assoc.

[189] S. Kapur, A. G. Phillips, and T. R. Insel, "Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?" *Molecular psychiatry*, vol. 17, no. 12, pp. 1174–1179, 2012, Publisher: Nature Publishing Group.

[190] A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer, "Forecasting the onset and course of mental illness with Twitter data," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017, Publisher: Nature Publishing Group.

[191] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009, Publisher: Elsevier.

[192] 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)*, IEEE, 2014, pp. 960–964.

[193] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *the Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002, Publisher: Acoustical Society of America.

[194] T. Hacki, "Klassifizierung von glottiscysfunktionen mit hilfe der elektroglottographie [classification of glottal dysfunctions on the basis of electroglottography]," *Folia phoniatrica*, vol. 41, no. 1, pp. 43–48, 1989, Publisher: Karger.

[195] J. Kane, C. Gobl, S. Scherer, and L.-P. Morency, "A comparative study of glottal open quotient estimation techniques," *BDL*, vol. 178, no. 15.17, pp. 0–41, 2013.

[196] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. C. Guiod, and S. L. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 6, pp. 1212–1223, 1995, Publisher: ASHA.

[197] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter—a new method for quantification of the glottal flow," *Speech Communication*, vol. 22, no. 1, pp. 67–79, 1997, Publisher: Elsevier.

[198] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1170–1179, 2013, Publisher: IEEE.

[199] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[200] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews.," in *Interspeech*, 2018, pp. 1716–1720.

[201] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[202] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[203] C. Shin, S.-H. Lee, K.-M. Han, H.-K. Yoon, and C. Han, "Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: Analysis of psychiatric outpatient data," *Psychiatry investigation*, vol. 16, no. 4, p. 300, 2019, Publisher: Korean Neuropsychiatric Association.

[204] I. Madhavi, S. Chamishka, R. Nawaratne, V. Nanayakkara, D. Alahakoon, and D. De Silva, "A Deep Learning Approach for Work Related Stress Detection from Audio Streams in Cyber Physical Environments," in *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, vol. 1, IEEE, 2020, pp. 929–936.

[205] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 45–51.

[206] N. Srimadhur and S Lalitha, "An end-to-end model for detection and assessment of depression levels using speech," *Procedia Computer Science*, vol. 171, pp. 12–21, 2020, Publisher: Elsevier.

[207] M. Niu, K. Chen, Q. Chen, and L. Yang, "HCAG: A Hierarchical Context-Aware Graph Attention Model for Depression Detection," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4235–4239. DOI: 10.1109/ICASSP39728.2021.9413486.

[208] M. Hiraga, "Predicting depression for japanese blog text," in *Proceedings of ACL 2017, Student Research Workshop*, 2017, pp. 107–113.

[209] S. Amir, M. Dredze, and J. W. Ayers, "Mental health surveillance over social media with digital cohorts," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 114–120.

[210] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, IEEE, 2013, pp. 1–8.

[211] G. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in Twitter," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[212] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2011, Publisher: IEEE.

[213] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 881–888.

[214] Z. Xu, V. Pérez-Rosas, and R. Mihalcea, "Inferring social media users' mental health status from multimodal information," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 6292–6299.

[215] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE transactions on affective computing*, vol. 4, no. 2, pp. 142–150, 2012, Publisher: IEEE.

[216] M. Valstar *et al.*, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 3–10.

[217] N. C. Maddage, R. Senaratne, L.-S. A. Low, M. Lech, and N. Allen, "Video-based detection of the clinical depression in adolescents," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2009, pp. 3723–3726.

[218] K. E. B. Ooi, L.-S. A. Low, M. Lech, and N. Allen, "Prediction of clinical depression in adolescents using facial image analaysis," in *WIAMIS 2011: 12th International Workshop on Image Analysis for Multimedia Interactive Services, Delft, The Netherlands, April 13-15, 2011*, Citeseer, 2011.

[219] K.-Y. Huang, C.-H. Wu, Y.-T. Kuo, H.-H. Yen, F.-L. Jang, and Y.-H. Chiu, "Data collection of elicited facial expressions and speech responses for mood disorder detection," in *2015 International Conference on Orange Technologies (ICOT)*, 2015, pp. 42–45. DOI: 10.1109/ICOT.2015.7498502.

[220] B. Zou *et al.*, "Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders," *IEEE Transactions on Affective Computing*, 2022, Publisher: IEEE.

[221] V. Mitra, A. Tsiartas, and E. Shriberg, "Noise and reverberation effects on depression detection from speech," in *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 5795–5799.

[222] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: A comparison between spontaneous and read speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 7547–7551.

[223] S. Park, S. Scherer, J. Gratch, P. J. Carnevale, and L.-P. Morency, "I can already guess your answer: Predicting respondent reactions during dyadic negotiation," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 86–96, 2015, Publisher: IEEE.

[224] D. Zhang, S. Li, Q. Zhu, and G. Zhou, "Modeling the clause-level structure to multimodal sentiment analysis via reinforcement learning," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2019, pp. 730–735.

[225] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression severity prediction based on biomarkers of psychomotor retardation," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 37–43.

[226] B. Sun *et al.*, "A random forest regression method with selected-text feature for depression assessment," in *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, 2017, pp. 61–68.

[227] D Freedman and P Diaconis, *On the histogram as a density estimator: L2 theory, Probab. Theory Rel., 57, 453–476*, 1981.

[228] K. Mao *et al.*, "Prediction of depression severity based on the prosodic and semantic features with bidirectional lstm and time distributed cnn," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2251–2265, 2023. DOI: 10.1109/TAFFC.2022.3154332.

[229] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007, Publisher: Elsevier.

[230] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2010, Publisher: IEEE.

[231] H. Jiang *et al.*, "Detecting depression using an ensemble logistic regression model based on multiple speech features," *Computational and mathematical methods in medicine*, vol. 2018, 2018, Publisher: Hindawi.

[232] S. Alghowinem *et al.*, "A comparative study of different classifiers for detecting depression from spontaneous speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 8022–8026.

[233] A. Pampouchidou *et al.*, "Depression assessment by fusing high and low level features from audio, video, and text," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 27–34.

[234] S. Dham, A. Sharma, and A. Dhall, "Depression scale recognition from audio, visual and text analysis," *arXiv preprint arXiv:1709.05865*, 2017.

[235] A. Samareh, Y. Jin, Z. Wang, X. Chang, and S. Huang, "Predicting depression severity by multi-modal feature engineering and fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Issue: 1, vol. 32, 2018.

[236] G. Coppersmith, K. Ngo, R. Leary, and A. Wood, "Exploratory analysis of social media prior to a suicide attempt," in *Proceedings of the third workshop on computational linguistics and clinical psychology*, 2016, pp. 106–117.

[237] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, "Discovering shifts to suicidal ideation from mental health content in social media," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 2098–2110.

[238] N. Jaques, S. Taylor, E. Nosakhare, A. Sano, and R. Picard, "Multi-task learning for predicting health, stress, and happiness," in *NIPS Workshop on Machine Learning for Healthcare*, 2016.

[239] K. Saha, B. Sugar, J. Torous, B. Abrahao, E. Kıcıman, and M. De Choudhury, "A social media study on the effects of psychiatric medication use," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, 2019, pp. 440–451.

[240] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.