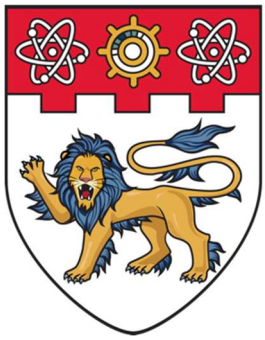


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Automated Socio-Cognitive Assessment of Patients with Schizophrenia and Depression

Xu Shihao

School of Electrical and Electronic Engineering

2022

Automated Socio-Cognitive Assessment of Patients with Schizophrenia and Depression

Xu Shihao

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2022

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

20-May-2022
.....
Date

.....
Xu Shihao

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

20-May-2022

.....

Date

.....

Prof. Andy Khong

Authorship Attribution Statement

This thesis contains material from 4 papers published in the following papers accepted at conferences in which I am listed as an author.

Parts of chapter 1, 2, 3, 4, and 5 are submitted as S. Xu, Z. Yang, D. Chakraborty, Y.H. Victoria Chua, J. Dauwels, T. Serenella, S. Winkler, M. Birnbaum, B. Tan, J. Lee, J. Danwels, "Identifying Psychiatric Manifestations in Schizophrenia and Depression from Audio-Visual Behavioural Indicators through a Machine-Learning Approach", npj schizophrenia, submitted.

The contributions of the co-authors are as follows:

- The entire project was conceptualized by the technical PI and my first supervisor, Assoc. Prof. Justin Dauwels who was affiliated with the Nanyang Technological University (NTU) and the clinical PI, Dr. Jimmy Lee Chee Keong from the Institute of Mental Health (IMH). Both of them also revised the manuscript.
- I collected all audio-visual recordings of the data and performed all the data preprocessing, feature extraction, and machine learning work at the School of Electrical and Electronic Engineering of NTU. I also prepared the manuscript drafts.
- Ms. Zixu Yang from the Institute of Mental Health (IMH) recruited the participants of the study, conducted interviews with them, assisted in collecting data, and revised the manuscript.
- Together with Dr. Chakraborty, I co-designed the non-verbal speech features and body-movement features.
- Assoc. Prof. Stefan Winkler shared Opsis API for emotion recognition, which is used in this study.
- Y.H. Victoria Chua, M. Birnbaum, and T. Serenella had roles in the result interpretation and revision of the manuscript.

Parts of chapter 1, 2, and 4 are published as S. Xu, Z. Yang, D. Chakraborty, Y.H. Victoria Chua, J. Dauwels, D. Thalmann, N.M. Thalmann, B. Tan, and J. Lee, "Automated Verbal and Non-verbal Speech Analysis of Interviews of Individuals with Schizophrenia and Depression", in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jul. 2019, pp. 225–228. doi: 10.1109/EMBC.2019.8857071.

and

S. Xu, Z. Yang, D. Chakraborty, Y. Tahir, T. Maszczyk, Y.H. Victoria Chua, J. Dauwels, D. Thalmann, N.M. Thalmann, B. Tan, and J. Lee, 'Automatic Verbal Analysis of Interviews with Schizophrenic Patients', in 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), Shanghai, China, Nov. 2018, pp. 1–5. doi: 10.1109/ICDSP.2018.8631830.

and

S. Xu, Z. Yang, D. Chakraborty, Y. Tahir, T. Maszczyk, Y.H. Victoria Chua, J. Dauwels, D. Thalmann, N.M. Thalmann, B. Tan, and J. Lee, "Automated Lexical Analysis of Interviews with Individuals with Schizophrenia", in 9th International Workshop on Spoken Dialogue System Technology, Singapore, 2019, pp. 185–197. doi: 10.1007/978-981-13-9443-0_16

The contributions of the co-authors are as follows:

- The entire project was conceptualized by the technical PI and my first supervisor, Assoc. Prof. Justin Dauwels who was aliated with the Nanyang Technological University (NTU) and the clinical PI, Dr. Jimmy Lee Chee Keong from the Institute of Mental Health (IMH). Both of them also revised the manuscript.
- I had the major responsibility of audio single processing, natural language processing, and machine learning. I also prepared the manuscript drafts.
- For these papers, Dr. Chakraborty and Dr. Tahir had roles in data acquisition and guided me with the extraction of the openSMILE and conversational features.
- Tomasz Maszczyk helped me with the use of the speech recognition toolkit.
- Professors Nadia Magnenat-Thalmann, Daniel Thalmann, and Bhing-Leet Tan provided initial direction to the project and helped us secure funding for the project. measurement data.
- The contributions of all other authors remained the same as the previous publication.

20-May-2022

Date

Xu Shihao

Acknowledgements

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

Firstly, I would like to express my sincere gratitude to my advisors Prof. Justin Dauwels and Prof. Andy Khong for the continuous support of my Ph.D. study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would like to acknowledge my colleagues, Debsubhra Chakraborty, Yasir Tahir, and Yi Han Victoria Chua, for their wonderful collaboration. I want to thank you for your patient support, help, and understanding.

I would also like to thank my mentor, Michel Birnbaum, for his valuable guidance and foresight throughout my studies. We work in a close-knit community that shares passion, beliefs, and values. The goal you set always steers us in the right direction, and taught me the importance of technology transfer.

I wish to thank the people in the Institute of Mental Health in Singapore, Mrs. Zixu Yang, and Dr. Jimmy Lee Chee Keong, for their cooperation throughout my studies. Thanks for providing me valuable clinical data and feedback for my publications.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me. Finally, I could not have completed this dissertation without the support of my friends, Xinyi Zhang, Yixin Wang, Hexin Liu, Siyao Yang, Hongquan Long, Qi Li, Guohao Peng, and Ajay Vishwanath, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

"Your time is limited, so don't waste it living someone else's life. Don't be trapped by dogma – which is living with the results of other people's thinking."

—Steve Jobs

Abstract

Schizophrenia and depression are two of the top 15 chronic mental disorders with severe impact on the people affected. There is still limited understanding in the medical community on whether machine learning methods applied to objective audio-visual behavioral cues can predict the severity of negative, cognitive, and general psychiatric symptoms and distinguish schizophrenia and depression patients from healthy controls with clinically relevant performance.

The objective of this thesis is to design and validate machine-learning pipelines to automatically distinguish schizophrenia patients and depression patients from healthy controls and to predict the severity of negative, cognitive, and general psychiatric symptoms. The key question attempting to answer is: can automated analysis of audio-visual signals predict the severity of negative, cognitive, and general psychiatric symptoms of schizophrenia and depression, and differentiate patients from healthy controls?

Specifically, this thesis analyzed the speech, facial expressions, and body movement recordings for schizophrenia and depression patients in two separate studies. The first study was conducted from 2014 to 2016, which included the recruitment of the cohort. The study was conducted with 58 patients with schizophrenia and 29 healthy controls over three sessions: at week 0, week 2, and week 12. The second study was conducted between 2017 and 2019 involving 50 patients with depression, 50 patients with schizophrenia, and 50 healthy control subjects, where only one session was conducted for each participant. In both studies, all subjects spoke English and were matched in age, gender, educational background, and ethnicity. Patients were then selected for persistent and predominantly negative symptoms with minimal positive symptoms. The baseline session of the first study was combined with the second study for model training and leave-one-out cross-validation, resulting in a total of 228 participants (103 patients with schizophrenia, 50 patients with depression, and 75 healthy controls), where 5 schizophrenia patients and 4 controls were excluded due to equipment malfunction or error in the consent form.

This thesis demonstrated the ability of machine learning algorithms using verbal, non-verbal, facial, and body-movement behavioral cues and signals to predict clinical assessment outcomes of overall negative symptoms, cognitive symptoms, general psychiatric symptoms, and scales related to diminished expression, and to classify schizophrenia, depression, and healthy groups. The results were obtained for recorded interviews of 103 patients with schizophrenia, 50 patients with depression, and 75 healthy controls. The proposed machine learning system achieves a moderate-high accuracy for classifying the total score of negative symptoms (balanced accuracy, 76.0%; sensitivity, 80.2%; specificity, 71.8%), the composite score of cognitive symptoms (balanced accuracy, 75.6%; sensitivity, 80.8%; specificity, 70.5%), and total score of general psychiatric symptoms (balanced accuracy, 73.6%; sensitivity, 83.3%; specificity, 63.8%). In particular, our results demonstrate the success of predicting assessment ratings that are directly or indirectly related to diminished expressions with a moderate-high balanced accuracy (>75%), such as restricted speech, affective blunting, and token motor test, while achieving relatively poor results (<65%) on semantic fluency and resistance factor scores, which are not directly related to diminished expression. Furthermore, the proposed system is able to differentiate schizophrenia and depression recordings from healthy control recordings with 82.3% balanced accuracy, differentiate between depression and schizophrenia with a balanced accuracy of 84.7%, and distinguish the three groups combined (schizophrenia, depression, and healthy controls) with a three-class classification accuracy of 68.7%.

These results suggest that, by extracting behavioral cues from audio and video clinical data, the proposed system is able to differentiate among three subject groups at a clinically relevant level of accuracy as well as to detect negative, cognitive, and general psychiatric symptoms at good clinical performance levels. These results were obtained on a Singapore-based cohort of Asian ethnicity. In the future, we will continue developing and optimizing the proposed system on a more diversified patient population, to explore cultural differences in the presentation of specific symptoms. These research efforts may eventually lead to digital phenotyping technologies for long-term monitoring of patients remotely, ultimately resulting in potentially better care for out-patient populations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Mental Illnesses: Insights and Challenges

Schizophrenia and depression are two of the top 15 chronic mental disorders with severe impacts on the people affected [6]. Depression affects nearly three hundred million people globally and is characterized by low mood, loss of interest or enjoyment, feeling guilty or low self-worth, loss of sleep or appetite, tiredness, and lack of concentration [7]. Schizophrenia affects more than 20 million people globally, and patients suffering from schizophrenia have about fifteen to twenty-five years of reduced life expectancy compared with the general population [6, 8]. Schizophrenia is characterized broadly by negative (e.g., anhedonia, asociality, avolition, affective blunting, and alogia), positive (e.g., delusions and hallucinations), and cognitive (e.g., attention, memory, and problem solving) symptoms [9]. Recent studies suggest that negative symptoms may not be unique to schizophrenia as previously thought, as those symptoms have been observed in people with depression and other mood disorders [10, 11]. Similarly, cognitive deficits in people with depression have become a clinically relevant target for treatment [12, 13]. Although positive symptoms are often easily identifiable and treatable with effective medications, negative and cognitive symptoms are often overlooked, are less responsive to pharmacological interventions [14], and are more closely associated with poor treatment response, poor functional outcomes, resulting in a diminished quality of life for schizophrenia patients [15, 16]. Moreover, patients themselves may not

be aware of the presence and impact of negative symptoms, highlighting a clear clinical need for effective assessment and monitoring of these symptoms [17].

In clinical practice today, the manuals for assessing and diagnosing mental disorders (e.g., DSM-5) and psychometric tools (e.g., PANSS) are considered the gold standard of diagnostic and assessment for mental illnesses. However, these tools rely on the rater's experience and intuition; consequently, they introduce a certain amount of subjectivity, are resource-intensive, and offer limited information concerning the temporal and spatial dynamics underlying clinical symptoms and manifestations [18]. Both clinical interview and assessment scales allow clinicians to assess patients over a limited period. However, since clinical assessments are only conducted occasionally, the scales cannot be monitored continuously over time. As a result, subtle changes in trends in these scales might be missed, and therefore it is more difficult to track the effectiveness of a treatment reliably. Clinicians have experienced difficulties in effectively identifying and treating negative symptoms during a short clinical visit. There is therefore a need to develop objective methods that allow diagnosis, assessment, and monitoring beyond healthcare settings across different time points [17].

Vocal and facial expressions are core components clinicians rely on during diagnosis and assessment and, in particular, speech impairment, one of the hallmark indicators for negative symptoms, cognitive impairments (e.g., disorganization of thoughts), and psychomotor abnormalities (e.g., excessive motor activity or catatonia). Non-verbal communicative behaviors are critically interlinked with the clinical manifestations of negative symptoms. Across various assessment and rating scales, clinicians outline manifestations such as reduced facial expressions, eye contact, gestures, body movements, and vocal expression to diagnose negative symptoms [7, 19]. For example, prolonged time to respond is one item within the 16-item Negative Symptom Assessment (NSA-16) for clinicians to assess in patients [20]. Similarly, the 5-factor conceptualization of negative symptoms by the National Institute of Mental Health (NIMH) highlights how expressive deficits manifest through verbal and non-verbal communication [10, 21], while experiential deficits may be exhibited in non-verbal behaviors, e.g., psychomotor retardation as a manifestation of avolition [10]. Studies have demonstrated a link between limited cognitive resources and expressive deficits in patients diagnosed with schizotypy [22]. The exhaustion of cognitive resources resulted in patients generating

fewer words, less semantic complex speech, and longer average pauses in speech [23]. Altogether, it is unsurprising that researchers have leveraged on a myriad of techniques over the past decade to mine and analyze non-verbal and verbal cues to assess negative and cognitive symptoms.

Digital phenotyping, defined as the moment-by-moment quantification of the individual level human phenotype in situ using data from personal digital devices [24], offers an innovative lens to observe behaviors in naturalistic and longitudinal settings [25]. This approach also fits naturally into the NIMH's Research Domain Criteria (RDoC) framework that suggests new ways of classifying mental disorders based on dimensions of observable behavior and neurobiological measures, potentially leading in the future to more effective diagnostic tools in psychiatry [26]. Several implementations of digital phenotyping have been designed, guided by the RDoC, to quantify behaviors associated with mental illnesses objectively [27, 28]. Along similar lines, many studies that analyze audio and visual data of schizophrenia patients have demonstrated abnormalities in language [29–31], speech [32–34], facial expressions [35–38], and motor [39–41] behaviors. Similar studies of depression patients have identified abnormalities in verbal [42–44] and non-verbal behaviors [45–49], facial expressions [47, 50–52], and body movement [53, 54] associated with depression. This stream of the literature suggests that digital phenotyping is a promising avenue towards objective behavioral measures for characterizing mental disorders.

Despite the exciting progress in individual-level digital phenotyping, our understanding of mapping digital phenotypes to symptoms is still limited. Recent findings suggest that mental illness patients' vocal and facial characteristics are associated with blunted affect and alogia [27, 55, 56]. Cohen et al. found that the clinical outcomes of blunted affect and alogia can be accurately predicted based on vocal features [56]. However, it remains unclear whether the behavioral phenotyping fueled by machine learning allows us to accurately predict the overall severity of negative symptoms and other psychiatric symptoms. To the best of our knowledge, machine learning pipelines for detecting cognitive symptoms for schizophrenia and depression have yet been developed. Moreover, except for Lott and Kliper [57, 58], none of the existing studies consider differential diagnosis; instead, they are limited to a single psychiatric disorder. Under RDoC, it has been postulated that many

different neuropsychiatric diseases share symptoms [59], hence it is crucial to explore the underlying mechanisms across various categorically defined disorders and different symptoms [60].

Clinical evaluation typically requires combining multiple heterogeneous sources of information, such as the verbal, non-verbal, and facial expressions of patients during interviews, various types of psychiatric instruments, and self-reporting and reports by family members/caregivers. In the past ten years, due to the release of several multi-sensor open-source data sets (such as DAIC, BackDog, and Pitt) [50], many studies utilized multimodal analysis and machine learning to diagnose depression and found that a combination of multiple modes usually outperformed using a single mode [47, 61, 62]. However, the potential of combining multiple modalities for diagnosis and measuring the psychiatric state of patients with schizophrenia has not been investigated so far. Furthermore, it is also not obvious whether the tandem of digital phenotyping and machine learning can easily be implemented in a lightweight mobile app for monitoring patients over a long period while maintaining a high level of accuracy and keeping the collected data secure, without jeopardizing the privacy of these at-risk patients. To achieve this vision, multiple psychiatric disorders, multiple types of symptoms, and multiples types of behavioral patterns were considered in this thesis simultaneously, which is aligned with the RDoC model, which seeks to establish correlates between domain and constructs on the one hand, and behaviors and other units of analysis on the other hand [63].

1.2 Social Signal Processing

Social Signal Processing (SSP) is a cross-disciplinary research domain that spans across several core concepts in psychology, anthropology, cognitive science, and digital signal processing [64]. During social interactions with colleagues, friends, parents, and strangers, we express and perceive the social signals to communicate effectively with others. The dynamics of social signals and their interpretations are not random but follow some principles and laws [65]. However, due to the complexity of human interactions and social expressions, it is challenging for a computer to identify those principles and laws that humans manage the use of the social.

From the perspective of social sciences, the fundamental issue of nature influences social signals interplay between the innate biological processes and social-culture processes [64]. On the one hand, some individuals may possess innate talents in social conversations and interactions. For example, such individuals may process a strong sense of empathy and are naturally attuned to emotions experienced by others. On the other hand, some social signals arise from an individual's environment and may differ across cultures and countries. For example, how people interpret and respond to negative emotions may vary across different cultures [66].

In SSP, researchers focus on social signals that are largely culture invariant. With reference to Figure 1.2, the social signals are embodied in the low-level behavioral cues, which represent a general modality of the interaction or communication. The low-level behavioral cues that a computer or a machine can detect could be separated into verbal and nonverbal cues [64]. For example, phenomena such as empathy, alogia, attention, and honesty are conveyed through a combination of multiple verbal and nonverbal cues. The verbal cues or linguistic cues include the words and sentences that the person used. The nonverbal cues include the physical appearance, gesture, eye contact, emotion, voice prosodic, communication distance, etc [67]. These social signals can be easily captured and interpreted during our daily life and underlie our understanding of human communication and relationships.

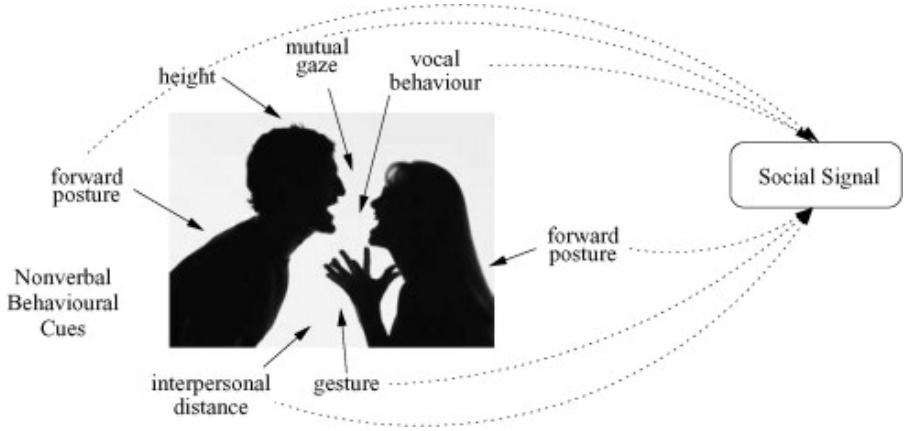


FIGURE 1.1: Behavior cues and social signals.

The study of nonverbal communication flourished in the early 1950s because semiotics aroused people's interest and was supported by recording technology [68]. Up to the 1970s, research was focused on extracting and understanding the different observations during human communications, particularly facial expressions

[69], gaze [70], posture [71], and gestures [72]. Extant literature on sign language found that each sign is represented by variant patterns in handshape, location, eye contact, and movement to convey a specific meaning or phonemes of a verbal language [73]. This research area highlighted the fundamental understanding of communication through a system of shared signals, e.g., language, and sought to understand how and what signals facilitate understanding during human communication. These findings were extended to the analysis of verbal communication, where researchers created extensive mappings between behavioral cues, how they are manifested (i.e., codes), and their functions (Fig 2).

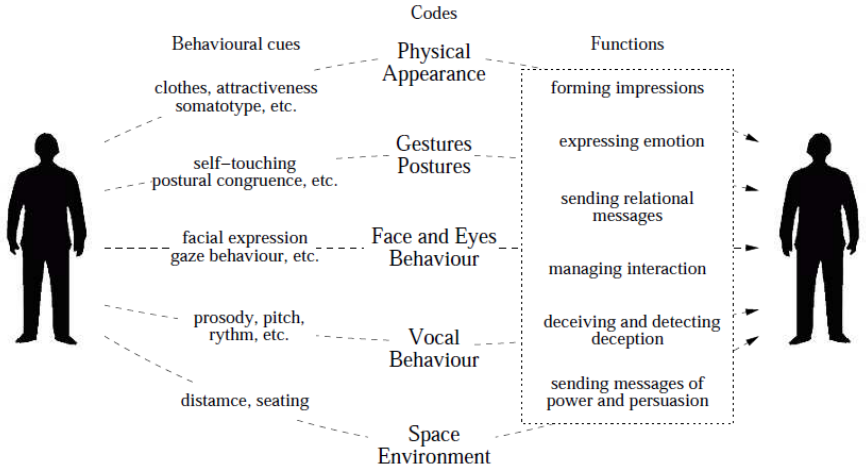


FIGURE 1.2: Codes and the behavioral cues.

1.2.1 Physical Appearance

Although people always say, "don't judge a book based on its cover", physical appearance does have a significant effect on our daily responses towards people, especially for first impressions [74]. The codes that affect social attractiveness include body characteristics such as height, body shape, skin, and hair color and contain artificial characteristics like dresses, makeup, perfume, jewelry, etc. The main findings in physical attractiveness are summarized in [75]: there is a strong correlation between a man's behavior and the person's physical appearance. For example, physical appearance positively correlates to the quality of social interaction when talking to a female and negatively correlates to that when talking to a male. However, for women, no such correlation exists; for both males and females,

assertiveness is positively correlated with the quantity and quality of social interactions; the more diverse attractiveness rating they received, the higher satisfaction they will feel; mutually initiated interactions happen more often with attractive men rather than unilateral interaction.

1.2.2 Gestures and Postures

Gestures are often used consciously and conventionalized, and it usually accompanies speech. Speakers may use different gestures in the same situation and, conversely, may use the same gestures in different situations [76]. For example, we often wave our hands to greet someone or bid goodbye. In addition, gestures are also used when we want to attract attention from the other party. However, postures are usually posed unconsciously. The distance between people, body orientation, and eye contact indicate what is happening during an interaction. These findings related to posture are summarized as follows [77]: the distance between the two speakers negatively correlates with the degree of enjoyment; humans adjust the amount of eye contact depending on how natural and comfortable they feel towards the person they are talking to. Interestingly, when feeling attracted to someone, eye contact may decrease slightly as well; for female communicators, there is a parabolic relationship between body orientation and attraction, similar to that of eye contact. For male, however, they only increase a few direct body orientations when they are conversing with a person he is fond of; the degree of asymmetry, openness, and types of position (e.g., arms folded, akimbo position) displayed by our arms and legs are indicators of attitude and degree of relaxation during an interaction. For instance, we are less likely to fold our arms across our chest when the other party is of high status and more likely to relax our limbs when talking to a person with lower status. Moreover, it is unlikely for one to adopt the arms-akimbo position when the other party is of high status. When talking to a person of lower status, body movements are more relaxed.

1.2.3 Facial Behavior

Facial expressions, emotion, and gaze behavior are the most reliable cues when understanding the social signals of a human. Many experiments have shown that

harmony judgments based solely on facial expressions achieve higher accuracy than that based solely on other behavioral cues [78].

The Facial Action Coding System (FACS) [79] is the most widely used protocol, developed by Paul Ekman and Wallace Friesen in 1978 and updated in 2002. In the beginning, Ekman and Friesen began electrically stimulating every muscle in the face and attempted to replicate these facial movements automatically. The smallest and most distinctive facial actions were defined as Facial Action Units. There are 27 such Facial Action Units on our entire face. Some researchers have also outlined some general guidelines about Action Units which typically co-occur and are frequently associated with the displays of emotions such as surprise, fear, happiness, sadness, disgust, and anger [80]. In addition, researchers also found a strong consensus of expressions of anger, fear, disgust, sadness, happiness, and contempt for people from many countries and cultures [81].

In the last ten years, automatic facial recognition and emotion recognition has become an increasingly popular research domain, with many applications. Many automatic facial expression analysis systems were designed to directly interpret the most basic facial emotions based on images and videos. For instance, Panic and Rothkrantz presented a prototype of an expert system for vision-based emotion recognition [82]. They proposed a dual-vision face model which includes 32 different facial actions and emotions. As a result, the overall emotional classification accuracy of the six basic emotional categories increases to about 91%. Moreover, Li Shan et al. reviewed the state-of-art datasets and algorithms leveraged to learn a discriminative representation of facial emotion recognition based on deep neural networks [83].

1.2.4 Vocal Behavior

The vocal-related cues include all spoken cues accompanying the verbal messages, such as linguistic features, prosody cues, articulation cues, phonetic cues, and conversational cues like interjection or interruption [84, 85].

Linguistic features include the usage of grammar, phase, and vocabulary in the speech or article. Differences in the use of language and words often reflect different psychological states and characteristics. Prosody cues are used to quantify the

variation in fundamental frequency, pitch, and timing accompanying the natural speech. Phonation is the vibration of the vocal folds to create sound. Articulation is the modification of the shape and position of the speech organs in the creation of sounds [86]. Finally, the conversational cues measure the turn-taking dialog pattern during the interaction, like how many times interjection happened and the average response time [87].

The Mel-frequency Cepstral Coefficients (MFCCs) [88] are widely used to represent sound. As spectral features are computed by taking the Fourier transform of the warped logarithmic spectrum, they contain information about rate changes in different spectrum bands. These MFCCs have been widely used in various speech and speaker recognition tasks [89].

1.2.5 Space and Environment Behavior

The physical distance between individuals usually corresponds to their social distance. However, many studies have shown that people tend to divide the space around them into circular zones where, by social standards, different kinds of people can enter [90]: For example, for family members, the distance can be less than 0.5 meters; for friends and colleagues, the comfortable distance is between 0.5 to 1.2 meters; and for formal relationships (such as bosses and employees) between 1.2 meters and 2.0 meters. These trends have been widely tested in Western societies but may change in other cultures.

1.2.6 Speech Adaptation Behavior

Numerous studies found that speakers spontaneously adapt, mimic, or converge with their interlocutors. Various studies describe this phenomenon differently, such as adaptation, [64], accommodation [91]. entrainment [5], alignment [92], etc. In [93], the words, mimicry in prosodic cues, have been used to represent the dynamics in conversational speech. Similarly, Spyros Kousidis et al. quantified acoustic convergence of recorded dialogues to perceive the spoken dialogue system is more "human-like" [94]. In our work, I followed the interaction adaptation theory (IAT) [95] and used "speech adaptation" to indicate the interpersonal behaviors on low-level speech patterns.

To date, various theories and models have been developed to understand the existence, causes, and effects of the speech adaptation phenomenon. Communication Accommodation Theory (CAT) [96] and IAT [95] are the latest two theories of centrality according to communication. CAT suggests people use communication to keep social distance from others. It uses convergence and divergence of verbal and nonverbal behaviors to indicate the speaker is adapting or not adapting to the interlocutor's speech [96]. The non-verbal cues explored in CAT include the length of utterance, interruptions, short and long pauses [97]; speaking rate and response time [98], and duration of pauses and switching pauses [99]. The above studies support the concepts that humans present different types of entrainment during conversational interaction. Speech adaptation seems like a complicated observation that different forms of entrainment may happen within the same dyad [96]. Besides, IAT defines speech adaptation as the degree of nonrandom, patterned, and synchronized in both timing and form [100]. It summarizes the predominant pattern of vocal entrainment as matching, complementarity, mirroring/similarity, reciprocity, compensation, convergence/divergence, synchrony, and maintenance/non-matching [101]. These forms of coordination and adaptation describe the verbal and nonverbal behaviors changing of one person according to the behaviors of its interlocutors in social interaction.

1.3 Motivation

Analysis performed in this thesis aims to design an automated and objective assessment platform for the symptom severity of schizophrenic and depressive patients. Schizophrenia is a long-term mental disease associated with language impairments that affect about one percent of the population. Depression is also a common and serious medical illness that negatively affects how you feel, think, and act. For example, depression causes feelings of sadness and/or a loss of interest in activities once enjoyed. Clinicians and trained professionals often observe through clinical interviews to diagnose and monitor mentally ill patients.

There are three typical symptoms of schizophrenia: positive symptoms, negative symptoms, and cognitive symptoms [102], and the negative symptom is also observed in people with depression [11]. Patients with positive symptoms may suffer

from hallucinations or delusions. They might hear, see, smell, or feel a thing differently from others. They often report hearing voices in their heads, and at times these voices may interact with each other. Negative symptoms indicate a loss of normal functioning and can include a diminished capacity to experience pleasure (anhedonia), decreased social affiliation (asociality), lack of motivation or drive (apathy), decreased outward expression of emotion (flat or blunted affect), and diminished speech (patients with depression often experience similar symptoms as well.). Cognitive symptoms are subtle and are often detected only when neuropsychological tests are performed. Patients with cognitive symptoms might present poor executive functioning (the ability to absorb and interpret the information and decide based on that information), inability to sustain attention, and problems with memories [103].

This thesis aims to assist psychiatrists in diagnosing and assessing mentally ill patients in a more objective and time-saving way. The current pain points in psychiatry that I try to overcome are:

- The popularity of psychotherapy. Current assessment procedures and treatments require high trained clinicians, which essentially limits the popularity of psychiatrists. The median number of mental health workers per 100,000 population is 9 [104].
- Not willing to seek help. Nearly two-thirds of them never seek treatment [105].
- Time-consuming. It is about 30 mins for the negative symptom assessment interview and 1 hour for cognitive symptom assessment. In addition, the median waiting time of a psychiatry appointment is about 50 days [106].
- Subjective and not accurate. Inter-rater reliability on total PANSS score ranged from 0.66 to 0.71 for tape observation and 0.92 to 0.99 for joint interviews [107].
- Cannot assess patients outside the hospital.
- Costly. In the US, a patient expects to pay about \$300 – \$500 for an initial consultation and \$100 - \$200 for the following session [108].

In order to diagnose and assess mental illness, trained professionals often conduct a mental assessment which includes an extensive repertoire of physical examinations, lab tests, and psychological evaluations. The defining symptoms of each mental illness are detailed in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [7], published by the American Psychiatric Association. A psychological evaluation is labor-intensive and time-consuming but necessary to form an accurate and comprehensive diagnosis and an appropriate treatment plan.

Current diagnosis and evaluation of schizophrenic patients with negative symptoms (and depressive patients) are costly, time-consuming, and inaccurate. This is because current assessment procedures and treatments require highly trained clinicians, which largely limit the popularity of psychotherapy. Moreover, the most common reasons that prevent people from obtaining the much needed mental health services are 1) fear and shame, 2) limited awareness, 3) feelings of inadequacy, 4) distrust, 5) unavailability, 6) high cost [109]. In many developing countries, psychological diagnosis and psychotherapy are still difficult to popularize. Hence, it is important to provide a more objective, time-saving, convenient, and affordable assessment platform to aid those who are willing to seek help.

Research performed in this thesis hypothesize that impairments in speech patterns highlighted by nonverbal and verbal cues are associated with higher severity of negative symptoms and poorer cognitive and functioning performance for schizophrenic and depressive when comparing to healthy controls. To test the above hypothesis, an automated speech processing pipeline was designed to analyze assessment interviews of mentally ill patients with trained clinicians. As a result, we found that it was possible to build a social signal processing system to analyze the behaviors of mentally ill patients in a more objective and convenient way. To this end, the overarching aim of this thesis is to develop a system that can automatically and objectively assess and diagnose the symptoms of mentally ill patients. Moreover, this thesis explores the mappings between speech cues and the severity of negative symptoms, cognitive impairments, and general psychiatric deficits in schizophrenia and depression.

To automatically analyze the speech cues of mental illnesses, speech signal processing, natural language processing, emotion recognition, and machine learning are the core tools. To develop machines that automatically capture and process the vocal and non-verbal signals, we could leverage and augment sensors to recognize

and process the vocal signals. However, the high-precision recognition and analysis of social signals for computers is still a challenge. On top of that, the machine also needs to balance the quality and computing time. Hence, optimizing the accuracy and performance of speech processing and automatic speech recognition is a significant challenge in my analysis. However, after extracting and recognizing the speech signals, machine learning algorithms are optimized and trained to learn and interpret patterns in the data, and the aim of automatically predicting the severity of new schizophrenic patients (unseen by the machine) can be achieved.

1.4 Contributions

This thesis evaluates the clinical utility of using automatic and objective data-driven approaches to differentiate diagnostic groups and predict the severity of various symptoms, and addresses some of the limitations of state-of-the-art digital phenotyping in psychiatry. Specifically, contributions of this thesis include:

- Multi-modal behavioral cues: thousands of behavioral cues are analyzed in this thesis, including low-level acoustic and prosodic cues (e.g., pitch and volume), linguistic cues, conversational measures, speech adaptation cues, motor cues, facial emotional expressions, and face and eye movements. Currently, as far as we know, there is no study trying to detect schizophrenia and predict the symptoms of schizophrenia using multiple modalities, and only a few for depression.

- Modular machine learning models for prediction and diagnosis: this thesis demonstrates that machine learning models can leverage this ocean of behavioral cues to detect various psychiatric symptoms and for differential diagnosis. The proposed machine learning pipeline is modular, in the sense that additional behavioral cues or other information about the subject (e.g., daily steps) can readily be integrated into the pipeline without needing to redesign the entire system.

- Multiple types of symptoms: this thesis explores the possibility of detecting a multitude of symptoms: negative, cognitive, and general psychiatric symptoms, as well as their subscales and domain scales, to facilitate the understanding for the clinician. To our best knowledge, there is no study found

to predict the cognitive symptoms using social signals, and just a few studies try to predict the negative symptoms and psychiatric symptoms automatically.

- Bivariate analysis of the speech for diagnosis and assessment: I proposed a quantitive method is proposed for measuring different forms of speech adaptation, which measures the adaptive behaviors of both patients and psychiatrists during the face-to-face clinical interview. As far as we know, there is no study evaluating the speech adaptation of schizophrenia patients, which is a common pattern that exists in human-human interaction.

- Across diseases: this thesis considers both depression and schizophrenia patients, besides healthy control subjects. Therefore, the severity of symptoms is predicted for the schizophrenia group and depression group, and the severity of symptoms across all groups is predicted.

- Multiple biomarkers: I discovered multiple digital biomarkers that differentiate patients from healthy controls on the dataset we collected. Some of biomarkers have not been observed in previous studies.

- Stability analysis: this thesis investigates whether the proposed digital phenotype models are consistent and stable across different recording sessions and multiple time points. This analysis constitutes the first small step towards long-term digital phenotyping on smartphones for longitudinal follow-up of psychiatric patients.

1.5 Outline of the Thesis

In this thesis, I extracted five different types of information from the audio and video recordings of interviews with patients with major depressive disorder and schizophrenia, and healthy controls: verbal and non-verbal audio cues, facial expressions, body movements, speech adaptation cues. The overall system pipeline is illustrated in Figure 1.3. I form these different types of cues and the analysis results in different chapters in this thesis.

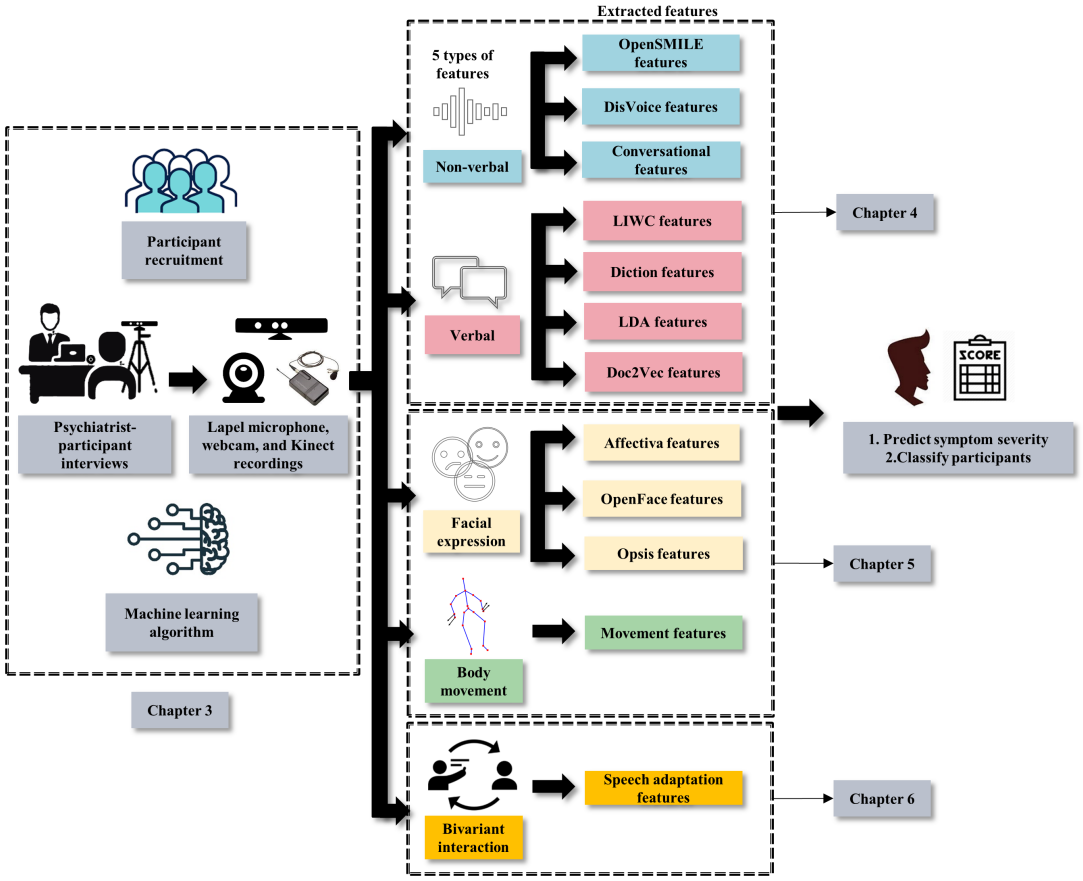


FIGURE 1.3: Diagram of the analysis pipeline.

Chapter 1 introduces the behavioral cues in the social signal processing domain, which is used for modeling mentally ill patients. This chapter also briefly discussed the recent insights and current challenges.

Chapter 2 reviews current machine learning research on schizophrenia and depression using audio-visual and motor behavioral cues.

Chapter 3 describes the system architecture of our approach, where details of the experiment, demographic information of the samples recruited, and the classification and feature ranking methods used in the analysis will be discussed.

Chapter 4, 5, and 6 demonstrate the process of extracting speech, facial expression and body movement, and the speech adaptation features, respectively. Numerical results of using these features in prediction and classification tasks in each chapter will be presented. Meanwhile, particulars of the verbal, non-verbal, facial, motor, and speech adaptation cues extracted from audio and video recordings will also be discussed.

Chapter 7 discusses the results and links our results to the literature, and Chapter 8 provides the concluding remarks, the limitation, and future view.

Chapter 2

Literature Review

Coupled with the speed and ease of data collection brought about by technological advances [110], several data-driven applications have greatly contributed to the diagnosis, interpretation, and understanding of mental illnesses [111]. In recent years, technological advances have largely decreased the time and effort to collect high-quality data. Many institutes have provided electronic medical records that involve a huge number of patients. At the same time, wearable devices and smartphones are becoming good choices for conducting long-term data collection to understand patients' behavior and social factors. However, with the abundance of data in the mental health field arises a new challenge: efficiently analyzing and interpreting this huge volume of data generated [112]. This work aims to keep up with the objective data-driven approach and develop a pipeline for automated analysis of behavioral content generated by schizophrenic and depressed patients. In the following sections, the candidate discusses the relevant research on schizophrenia and depression through automated and objective measures.

2.1 Verbal Analysis

Substantial advances in artificial intelligence and machine learning present promising avenues to develop objective clinical tools to aid clinicians. Linguistic analysis of content generated by psychiatric patients has become a popular mode of investigation these recent years. Text analysis programs such as the Linguistic Inquiry

and Word Count (LIWC) [113] and Diction [114] are often utilized for linguistic analysis of spoken and written content of schizophrenic patients, ranging from autobiographical narratives [115] and self-described transcriptions [116] to semi-structured and structured interviews [117, 118]. These dictionary-based methods measure word frequency in predefined categories, where the analysis pipeline is shown in Figure 2.1.

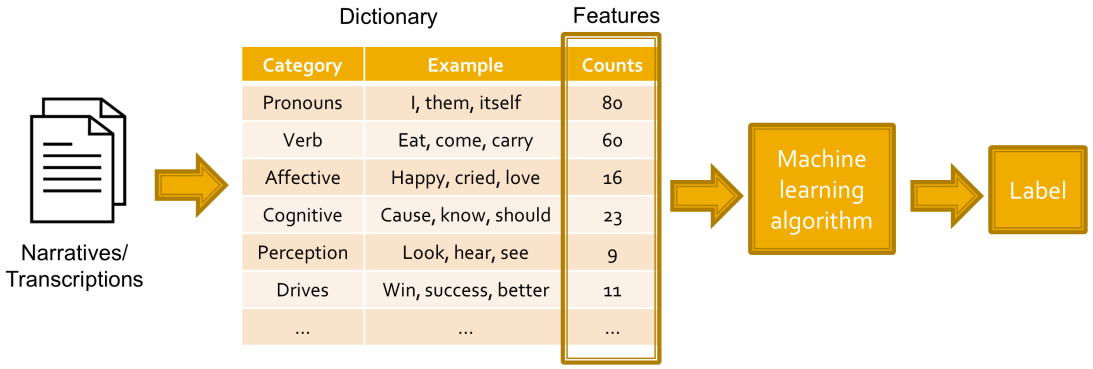


FIGURE 2.1: The classification pipeline of dictionary-based methods.

Kei Hong and his colleagues applied LIWC 2007 and Diction 6.0 to analyze the linguistic features of autobiographical narratives that differentiate schizophrenic patients and controls in [115]. In their experiment, they introduced a dataset of autobiographical narratives collected from 23 schizophrenic patients and 16 healthy controls. Each participant was asked to narrate their daily experiences of five different emotion states: happy, sad, anger, fear, and disgust. Since some participants told more than one story, there are 120 narratives from patients and 81 narratives from healthy controls using to classify schizophrenic patients and healthy controls, where the narratives are manually converted from audio recordings by a transcriber. They applied four classes of lexical features in their study: generic features (type-token ratio, mean word length, the average number of words per sentence, etc.), word identity features (the frequency of words in the narrative), dictionary-based features (features extracted from LIWC 2007 and Diction 6.0), and language model features (track probabilities of word sequences). After extracting the lexical features, they formed an automated feature selection and classification model with leave-one-subject-out cross-validation to predict clinical status (schizophrenia vs. control). They subsequently applied an SVM classifier to weight average the multiple prediction scores from five emotional-based narratives. They found distinct

differences in words related to LIWC category 'I' and Diction feature 'self' between patients and healthy controls. In addition, they found that patients were more likely to talk about the topics such as money, trouble, and family. The highest classification accuracy achieved to 74.4% at the best p-value cutoff, where they decided to select the features with p-values smaller than 0.0007.

In [116], Annie St-Hilaire et al. investigated the difference in linguistic usage and emotional expression between 48 diagnosed schizophrenic patients and 48 healthy controls. All participants underwent a structured diagnostic interview conducted by trained research assistants on two testing sessions. The Brief Psychiatric Rating Scale (BPRS) [119] was utilized to assess the symptoms of schizophrenic patients. During the interview, all participants completed a 10-minute conversational speech sample, where they were asked to describe themselves, their interests, hobbies, and activities. The speech samples were audiotaped and then transcribed and proofread manually. LIWC 2003 was used to analyze the text files and produce the word counting output. Using statistical testing methods, they found that patients' feelings were significantly more stressed than the feelings of controls during the self-description task. In addition, gender did not significantly influence the use of negative emotion words during the self-description task.

Another study by Minor [118] also explored the feasibility of utilizing lexical features to analyze the speech of schizophrenia automatically. In their research, 17 diagnosed schizophrenic patients and 29 schizoaffective disordered patients were involved in answering the Indiana Psychiatric Illness Interview's open-ended questions, which aimed to assess the perceptions of life and illness. The symptoms of schizophrenia were measured by the Positive and Negative Syndrome Scale (PANSS) [120]. The recordings were manually transcribed and processed for lexical analysis using LIWC 2007. They focused on the categories of LIWC related to emotion, cognitive, perception, biological, relativity, work, achievement, and social words. They found that the number of angry words and negative emotion words used in the interviews can significantly predict the severity of overall symptoms. In addition, reality distortion symptoms were linked with anger words, disorganized symptoms were associated with fewer words about work, and negative symptoms were associated with few social words. However, in this study, they did not apply machine learning algorithms to classify the patients with different severity of symptoms automatically.

Moreover, Parola et al. [121] conducted a multimodal assessment of the communicative pragmatic ability of patients with schizophrenia and healthy controls, such as communicative phenomena and expressive modalities. They recruited 32 patients with schizophrenia and 35 healthy controls, and all of them spoke native Italian. They then applied a decision tree classifier to distinguish these two classes and achieved an accuracy of 82%. They found that linguistic irony was the most relevant pragmatic phenomenon in distinguishing between the two groups.

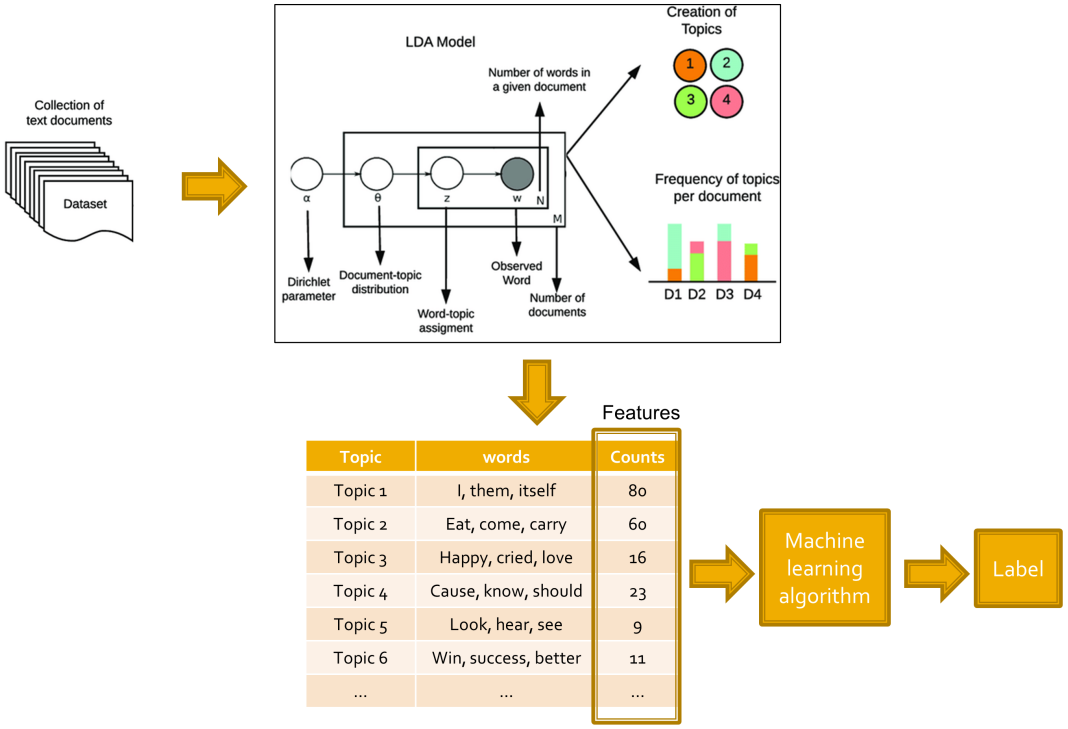


FIGURE 2.2: The classification pipeline of topic-based methods (e.g., LDA).

Notably, these studies have found significant differences in linguistic usage and conversational topics of schizophrenic patients, demonstrating the feasibility of using linguistic categories as features to analyze schizophrenic patients. Furthermore, automated linguistic analysis of speech impairments related to schizophrenia also employs context-based methods like Latent Semantic Analysis (LSA) [122], Latent Dirichlet Allocation (LDA) [123], Document to Vector (Doc2Vec) model [124], and Bidirectional Encoder Representations from Transformers (BERT) [125] to find subtle speech differences of schizophrenia. Specifically, the topic modeling algorithms, like LSA and LDA, extract the topics for a collection of documents and

the words for each topic automatically without the need for human-defined dictionaries, as shown in Figure 2.2. Moreover, the document embedding method, such as Doc2Vec, takes word order into account, generalizes to longer documents, and can learn from unlabelled data. In addition, BERT includes attention mechanisms to capture the context between words and between sentences from a large corpus. The architecture of using Doc2Vec and pre-trained BERT to classify documents is illustrated in Figure 2.3 and Figure 2.4.

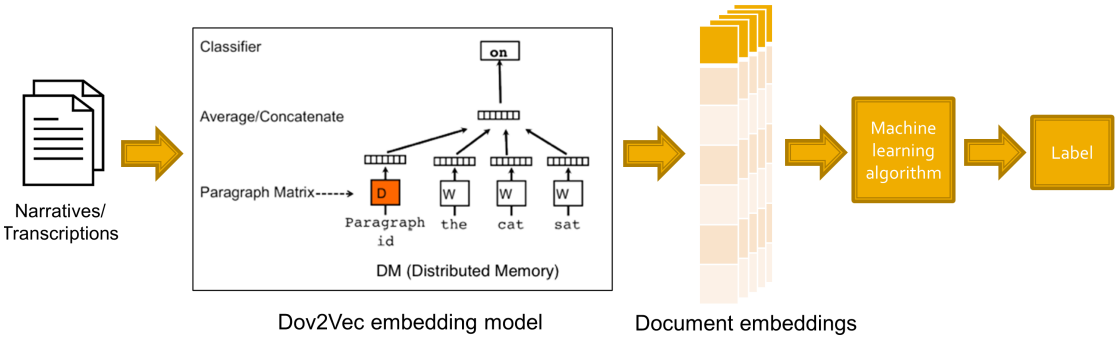


FIGURE 2.3: The classification pipeline of the document representation method (e.g., Doc2Vec).

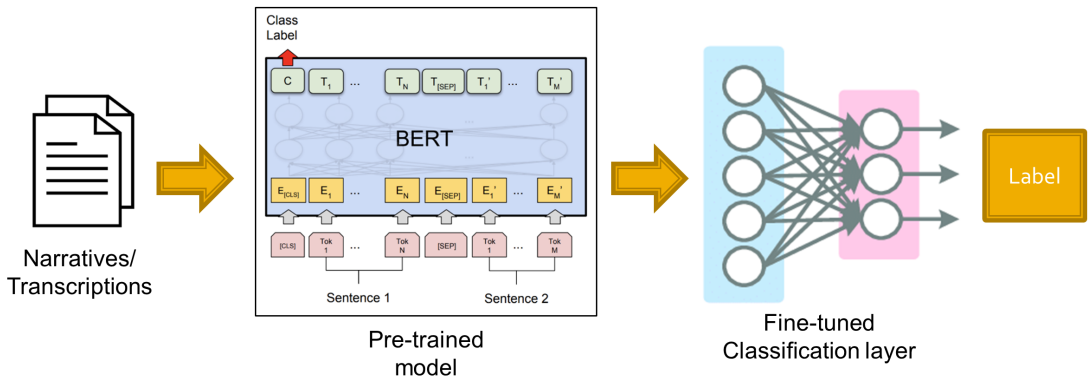


FIGURE 2.4: The classification pipeline of the fine-tuning word representation method (e.g., BERT).

In [126], Corcoran and colleagues attempted to distinguish schizophrenic patients from healthy controls and predict the onset of psychosis in high-risk youths. There were 93 clinical high-risk individuals, 21 healthy controls, and 16 first-episode psychosis in this study. First, the speech was recorded and manually transcribed when each participant took part in Caplan's "Story Game". Latent Semantic Analysis

(LSA) was then used to generate a series of semantic vectors, representing the frequency and semantic similarity of each word. Next, they formed a machine learning classification with cross-site validation to identify the healthy controls and recent-onset psychosis patients, and the classification accuracy is about 72%.

Bar et al. [127] also aimed to distinguish healthy controls from patients with schizophrenia using semantic features. There were 24 schizophrenia patients and 27 healthy controls in this study. They asked each participant 18 clinical questions and then manually transcribed answers into text. Next, they used fastText embeddings to represent the meaning of words and calculated derailment scores and coherence scores. Finally, they used these features to classify patients and healthy controls and achieved an accuracy of 81.5%. In [128], Tang et al. explored the linguistic characteristics of 20 patients with schizophrenia spectrum disorders (SSD) on three levels: individual words, parts-of-speech, and sentence-level coherence. Specifically, they utilized BERT, an embedding algorithm to incorporate bidirectional contexts. They found SSD used first-person singular pronouns more frequently than healthy controls. Besides, they leveraged the sentence embedding distance to represent the sentence-level coherence and found increased tangentiality among SSD than healthy controls. The classification accuracy with leave-one-out cross-validation achieved 87% using the Naive Bayes classifier.

One study reported the innovative method of using an automatic conversation topic modeling algorithm, Latent Dirichlet Allocation (LDA), to predict therapy outcomes with an accuracy of 75% [129]. Other context-based linguistic analysis methods (i.e., bag-of-words model, term frequency-inverse document frequency, and Doc2Vec) have similarly been utilized by researchers to differentiate patients of other mental illnesses such as autism spectrum disorder and healthy controls [130].

People with depression often suffer from negative life events and social defeat stresses [131, 132]. The abnormal language use of major depressive disorders is different from the people in a euthymic state of normal sadness [133]. In the cognitive theory of depression, focusing on the self, personal features, and the world are three main patterns [134]. Some studies found the frequency of using first-person singular pronouns and words with negative emotion (e.g., the words in the *sadness* category) were positively correlated with the depressive level [45, 135]. Bernard et al. [136] also suggested that depression affects the use of first-person

pronouns. These abnormalities may reveal the absence of positive thinking and the self-focused tendency of depression patients [44, 137].

One study analyzed the speech outcome of depressive, schizophrenic, and bipolar patients [57]. During the experiment, they asked emotionally neutral questions to condition stable patients and recorded their monologues. These speeches were then manually transcribed into texts by a professional linguist. The linguist assessed several abnormal language variables, including poor speech, speech pressure, loss of goals, illogical, and using these characteristics, and they obtained an overall performance of 72.7% correctly classified patients.

In [138], Choudhury et al. leveraged language and social behavior cues to diagnose major depression disorders in social media. They collected a group of social media posts from Twitter users diagnosed with depression and compared their behaviors before and after the diagnosis. They extracted the social engagement, linguistic, and emotional features and achieved an average accuracy of 70% in predicting the onset of depression in individuals. Their findings showcased that individuals with depression had slower social activity, greater negative emotion, and higher self-focusing. Similar findings were observed in the study of Wolohan as well [139], where they found that depressed social media users emphasize themselves more heavily than the control groups.

Trifu et al. also explored the evidence-based language indicators for major depressive disorders [44]. They analyzed the narrative speech of 75 depression patients and 42 healthy controls. They found they used mainly first-person pronouns and preferred singular pronouns. This study also observed that depressive patients used more past-tense words and spoke repetitive words more frequently. In conclusion, they linked these language changes to deficits in working memory, setting changes, strategic planning, attention, and psychomotor speed. The frequent use of first-person pronouns was also studied by Zimmermann [43]. They reported that self-attention words were correlated with the severity of depressive symptoms during a baseline session and a subsequent session.

In [137], Sonnenschein et al. compared the language use of depression patients with anxiety patients using LIWC. They analyzed the speech transcription of Cognitive Behavioral Therapy (CBT) for 85 outpatients, where 27 participants are depression patients but no anxiety disorder and 24 people had anxiety disorder but no

depression. They found depression patients used more sad words than anxiety patients. However, no differences were found in the use of first-person pronouns between them.

Smirnova et al. [133] studied about 400 written reports of patients with mild depression, people with normal sadness, and healthy controls. They observed the verbal usage of lexical repetition, omission of words, and verbs in continuous and present tenses were significant between people with normal sadness and healthy controls. Meanwhile, the notable language cues in discrimination of mild depression and healthy controls included analytic style, atypical word order, increased use of personal and indefinite pronouns, and verb use in continuous/imperfective and past tenses.

Arevian et al. conducted a longitudinal study on verbal information for 47 psychiatric patients diagnosed with schizophrenia, depression, and bipolar disorder. [140] They recorded about 1100 phone calls and 117 hours of speech and manually transcribed them into texts. Next, they found that the proportion of positive/negative emotion words, religious words, and sadness words were correlated with the global assessment score. Interestingly, in another study, Qureshi et al. proposed different attention-based neural networks to regress and classify both depression level and emotional intensity [141]. They utilized a universal sentence encoder to convert texts to vectors and trained the neural network to predict depression levels with learning emotion intensity at the same time.

2.2 Non-Verbal Analysis

Apart from the linguistic analysis of speech content from schizophrenia and depression patients, other studies have focused on the acoustic and non-verbal speech analysis of schizophrenia speech. Atypical voice patterns in schizophrenia are associated with clinical symptoms such as blunting of effect and may be an important indicator and contributor to the social impairments.

A recent review on acoustic patterns in schizophrenic speech found that patients exhibited reduced spoken time, pitch variability, and pause duration [32], pointing towards the possibility of identifying acoustic markers of schizophrenia. In this

review report, they contacted 37 authors and 54 studies to obtain the individual-level dataset. They then reviewed the literature quantifying acoustic patterns in schizophrenia and performed a meta-analysis of the evidence. Within the available datasets, they found the percentage of spoken time is a significant feature in 5 datasets, the proportion of spoken time is important in 5 datasets, and no effects were found for pause duration (7 datasets), speech rate (9 datasets), speech duration (5 datasets) and pitch intensity (5 datasets).

Rapcan and his team applied acoustic analysis to digital recordings of schizophrenic patients reading aloud [142]. There were 39 schizophrenic patients and 18 healthy controls who were required to read aloud a children's story. First, the speech data was collected in a quiet room with high quality. After which, 9 conversational features (number of pauses, mean pause duration, the proportion of silence, total recording time, energy, pitch, etc.) were computed from the audio signal. A classifier based on Linear Discriminant Analysis (LDA) with 18-fold cross-validation was then applied to differentiate patients and controls, and a classification accuracy of 79% was achieved. In addition, they found the pause-related features showed the most significant differences between patients with schizophrenia and healthy controls.

Several studies have applied automatic non-verbal conversational analysis to interviews with schizophrenic patients and found that non-verbal conversational cues such as mutual silence, response time, and natural turn-taking reliably distinguish patients and controls with an accuracy of 93%, and predict NSA-16 ratings with an accuracy of 80% [143], but do not reliably predict adherence to therapy [129].

Given the holistic nature of clinical assessment, the speech of schizophrenic patients is manually assessed by trained clinicians according to its semantic content, syntactic coherence, and conversational rapport with the interviewer [144]. Therefore, combining speech, linguistic, conversational, and acoustic analysis of schizophrenic patients presents an important direction towards developing objective tools to aid clinician diagnosis and assessment.

Cohen et al. [145] emphasized that accurate assessment of negative symptoms in patients with schizophrenia plays an important role in understanding and assessing schizophrenia. However, current assessment methods rely on subjective evaluation criteria, which are often time-consuming and inaccurate. Therefore, in their study,

natural audio recordings were collected and analyzed for 60 schizophrenic patients and 19 healthy controls in the semi-structured interview, during which the Scale for the Assessment of Positive Symptoms (SAPS) and the Scale for the Assessment of Negative Symptoms (SANS) was used to measure the positive symptoms and negative symptoms of each participant. Next, non-verbal features (inflection and speech rate) extracted from the speech signal and verbal features (the percentage of positive emotion words and negative emotion words) extracted from manual transcriptions by LIWC were used to explore the differences between computerized and traditional clinical-based measures. Their statistical findings provide an important understanding of the negative symptoms such as vocal blunt affect and alogia in both computer-based measurement and clinical assessment.

Moreover, from a Research Domain Criteria (RDoC) perspective, Cohen et al. analyzed the non-verbal speech features of 48 outpatients with stable schizophrenia and mood disorders [27]. They found negative symptoms, psychosis symptoms, and social functioning associated with decreased number of utterances and increased pauses. Moreover, they identified independent acoustic variables across 5 different studies, evaluated the impact of demographics (e.g., age, gender, and ethnicity), and examined the correlation between the vocal features and symptom severity [146]. They observed vocal variables are correlated to a series of psychiatric symptoms, especially to negative symptoms. However, after controlling for demographic and background factors, there was no significant difference in non-verbal features between the patient and the control group. In [55], Cohen et al. further analyzed about 800 video recordings. They observed that some ambulatory audiovisual features, e.g., pitch, jitter, and positive/negative facial expressions, are correlated with the blunted affect and alogia of patients with schizophrenia and schizoaffective disorder. Furthermore, they developed a machine-learning-based model to predict the assessment outputs of alogia and blunted affect for schizophrenia patients [56]. By using the non-verbal acoustic features extracted from picture talks and free speeches, they achieved an accuracy (balanced accuracy) of 85.0% (78.5%) and 92% (81%) when classifying different severity of blunted affect and alogia.

As described above, except for [34, 117], research on acoustic characteristics of schizophrenia mainly focuses on the pitch, intensity, pause, etc. Similar to depression, several studies revealed a marked decrease in pitch mean and pitch range

within speech produced by patients diagnosed with depression [33, 147], corresponding to the monotony of speech often clinically observed. Meanwhile, other acoustic features such as jitter, shimmer, and pitch variability tend to increase with the severity of depressive symptoms [33, 147, 148].

Furthermore, Cummins et al. conducted a comprehensive review of the speech-based research and characteristics of depression detection and suicide detection [49]. The change of non-verbal characteristics, including prosodic, source of the formant, and spectral features, are considered key digital biomarkers for depression detection. In the following, the main findings are summarized in these four aspects. For prosodic features pertaining to the monotonous and dull descriptions of speech affected by depression, it has been reported that the reduction of the average and range of F0 is positively correlated with the severity of depressive symptoms [149, 150]. As a result, the monotonous speech could be a result of the effects of psychomotor retardation [151]. Source features capture the information of voice production. In [148], they found that jitter and shimmer were significantly correlated with depression severity. Honig et al. [149] also observed a negative correlation between shimmer and the depressive level. Besides, abnormalities in formant-related features associated with depression patients may be responsible for changes in vocalization caused by muscle tension and mucus secretion; however, the observation on formant features seems inconsistent across different studies [150, 152, 153]. Lastly, the spectral analysis leveraged the high-level representations, e.g., MFCC, LSF, and power spectral density, to capture the frequency distribution of the speech signal at a specific time instance. The feature extraction pipeline of MFCC is shown in Figure 2.5. As a result, these high-level coefficients are able to predict the level of depression [154–157].

In recent years, advances in signal processing research have enabled researchers to use computerized speech features to study assessment or depression detection. The release of multiple multi-sensor open-source datasets for depression assessment (e.g., DAIC [158], BlackDog [159], and Pitt [160]) and the Audio/Visual Emotion Challenge and Workshop (AVEC) competition [161, 162] has aroused the interest of researchers. Alghowinem et al. performed a non-verbal analysis of the speech of depression patients on these three datasets [163]. They found that cross-validating samples from all three datasets provided the best depression detection results. However, their results also show that the classification accuracy is significantly

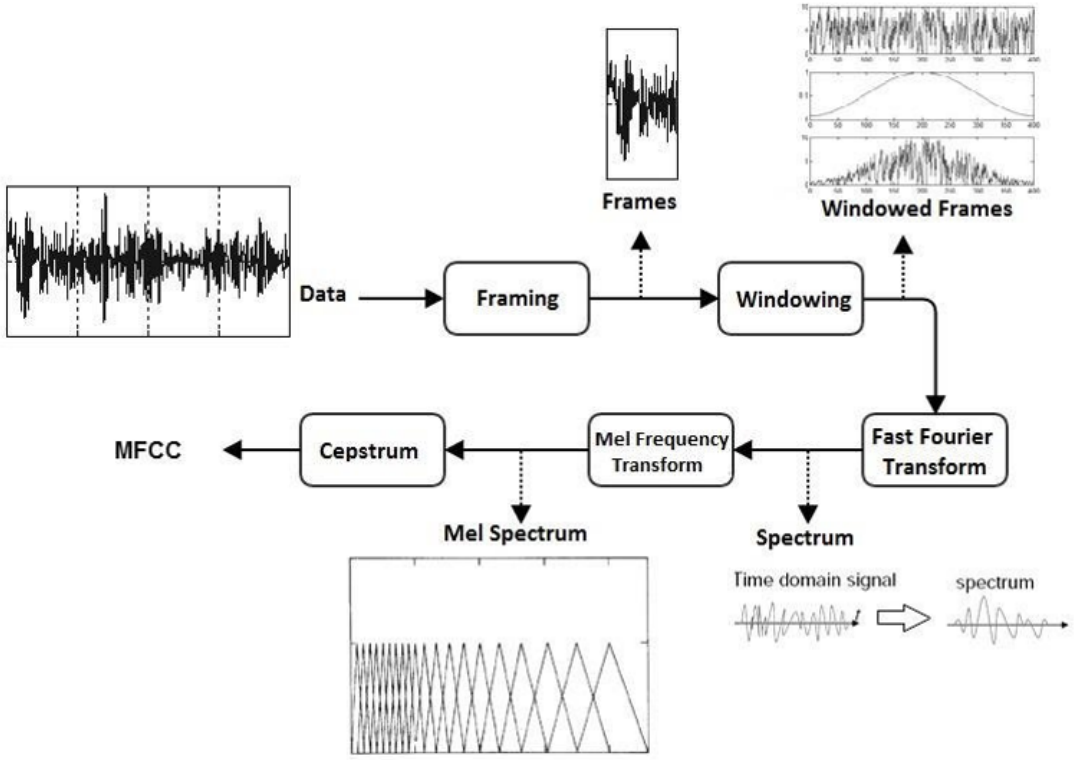


FIGURE 2.5: The extraction pipeline of the Mel-frequency Cepstral Coefficient (MFCC) [1].

high on the Pitt dataset (94.7%) but relatively worse on the AVEC dataset (68.8%). Besides, the authors also made a comprehensive review of depression detection from the perspective of feature selection [61]. They proposed a framework to generate feature weights from different feature selection methods. These feature weights explain the ability to distinguish the severity of depression. They found that speech and interactive functions are the strongest indicators in the detection of depression. Specifically, important voice features include F0, HNR, formant, and MFCC. Eye gaze direction and head movement are also important for depression detection.

The DAIC dataset contains audio and video recordings of the communication between participants with mood disorders and a virtual human [158]. These participants were marked with the severity of depression and PTSD. Stratou et al. [164] leveraged the Naive Bayes classifier with leave-one-participant-out cross-validation to identify depressed and non-depressed participants based on the features related to emotional expressions, action units, and head gestures. They observed that gender-dependent machine learning models achieved better performance than gender-independent models on depression detection. Besides, tree-based algorithms

were implemented to fuse audio-based, video-based, and text-based features to detect depression severity [165, 166]. In the study by Lam et al. [2], they implemented a context-aware analysis on the DAIC dataset using the transformer and 1D convolutional neural network (CNN) for end-to-end acoustic feature modeling, where the architecture is presented in Figure 2.6, and another study [157] applied dilated CNNs and investigated domain adaptation across the different corpus, aiming to improve the depression detection performance. CNN models are able to capture the smaller and simpler patterns through the convolution layer and select the important features through the pooling layer.

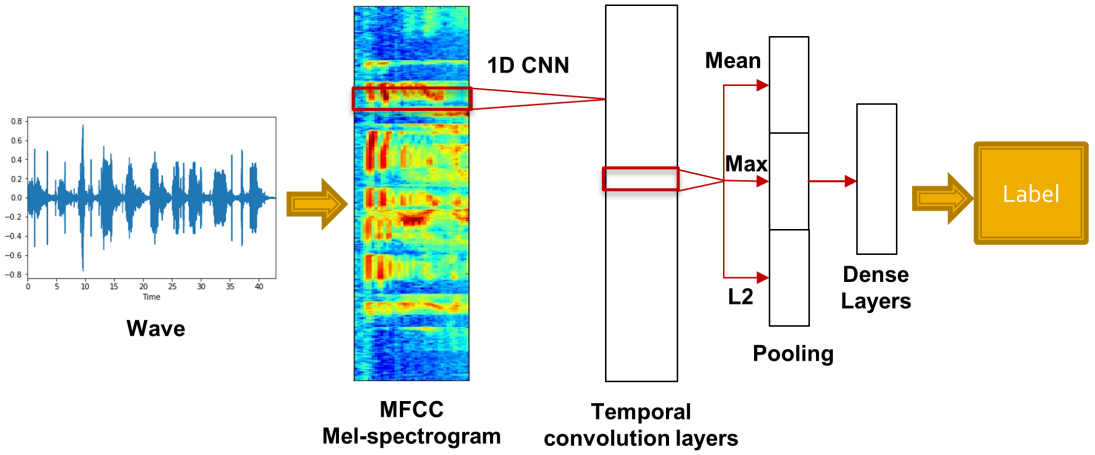


FIGURE 2.6: The classification pipeline of a CNN-based model for Mel-spectrogram [2].

The BlackDog dataset was collected in a continuous experiment at the BlackDog Institute in Sydney, Australia [159]. All participants with depression are met the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders - fourth edition) criteria and assessed by the Quick Inventory of Depressive Symptomatology self-report (QIDS-SR) [167]. In the study by Cummins [154], they utilized the low-level speech features to classify depressed/neutral speech (reading) of 47 speakers. They found using a combination of features related to MFCC and formant and a GMM-based classifier obtained an accuracy of 80%. Subsequently, they observed that the spectral variability was negatively correlated with the degree of depression [155], and Alghowinem et al. found that acoustic features, e.g., jitter, shimmer, energy, and loudness features, can distinguish depressed or neutral spontaneous speech through SVM [168, 169]. Moreover, Stasak et al. extracted verbal and non-verbal cues from reading speech in the BlackDog dataset and explored the differences in

emotional language features, hesitation, and speech errors between depressed and non-depressed people [170].

The audio and video recordings of the Pitt dataset are derived from 47 participants in a clinical trial of depression treatment conducted by the University of Pittsburgh Medical Center [160]. All patients were diagnosed with MDD by DSM-IV criteria and assessed by Hamilton Rating Scale for Depression (HRSD) [171]. In [160], the authors analyzed the vocal timing and fundamental frequency (F0) of both depressed patients and interviewers. They achieved an accuracy of about 69% for detecting the depression severity and found that vocal prosody accounted for about 60% of the variation in depression scores. Besides, they found the mean and variability of interviewers showed a strong correlation with the depression severity of participants, where the interviewer became more expressive (F0 is higher and less variable) when participants were more severe.

In addition, several studies analyzed the speech of depression patients along with other diseases. For example, Albuquerque et al. [172] evaluated the association between speech and both anxiety and depression symptoms. Patients with depressive symptoms showed a more significant relationship with acoustic parameters than those with anxiety symptoms. For example, adults with increased depressive symptoms showed longer vowel duration, longer total pause duration, and shorter total speech duration. Espinola et al. leveraged the SVM with PUK kernel and vocal features extracted from spontaneous speech to classify 22 MDDs and 11 healthy controls and provided an accuracy of 89.14% to detect MDD [173]. The non-verbal features, e.g., speech rate, pause time, and response time, of MDD, bipolar disorders, and healthy controls were also studied by Yamamoto [46]. They reported that depressed patients spoke slower, paused longer, had a longer response time than healthy controls, and had a longer response time than bipolar disorders.

2.3 Facial Expression

In addition to speech barriers, people with schizophrenia also suffer from deficits in encoding and decoding facial emotions. As summarized in [174], many studies examined the emotional perception and facial expression of patients with schizophrenia. In particular, for facial expression, patients with schizophrenia show a lower

proportion of positive expressions [175], especially anticipatory pleasure [176], and display more negative emotions than positive emotions [177]. In addition, Tremeau summarized a large number of publications [178]. They found that compared with normal people, schizophrenic patients tend to exhibit more negative emotions in real life and similar happiness and a higher degree of unhappiness in the evocative experiment. These emotional deficits, also referred to as blunted facial affect and anhedonia, are a prominent symptom of schizophrenia and may also appear in people treated for major depressive disorder [179, 180]. In most data-driven facial expression studies, researchers are based on FACS, which classifies human facial appearance based on facial movements [79]. Using FACS, the human facial expression can be encoded into dozens of Action Units (AUs), where each AU measures the movement of muscles or a group of muscles on the face, as illustrated in Figure 2.7. Furthermore, FACS is also associated with different emotions by means of the Emotional Facial Action Coding System (EMFACS) [181] and Facial Action Coding System Affect Interpretation Dictionary (FACSAID) [182].

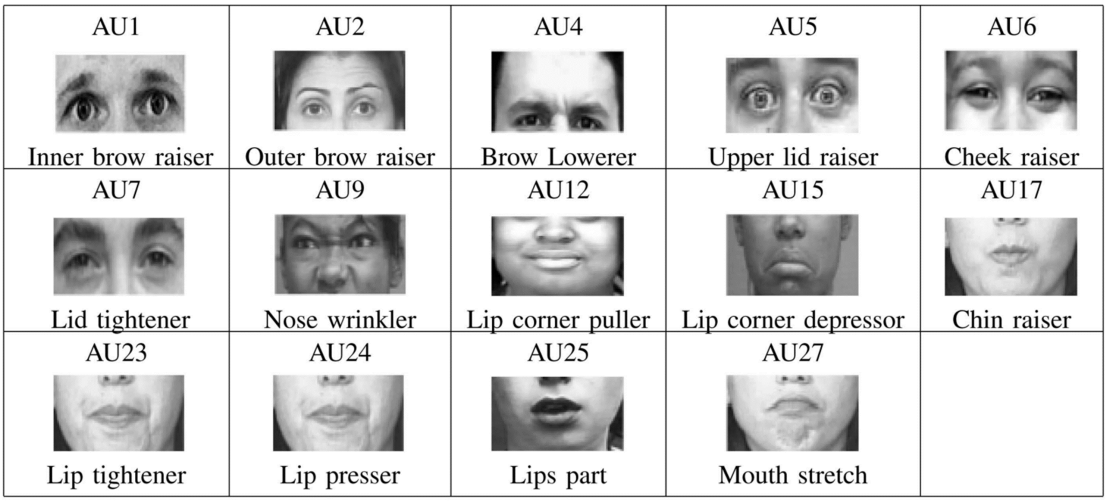


FIGURE 2.7: Illustration of facial action units [3].

In the study by Alvino et al. [183], they analyzed facial expression images of 12 patients with schizophrenia and 12 healthy controls and classified four common facial emotions (happy, sad, fear, and anger). They reported that the classification probability output between healthy controls and schizophrenia patients was significantly different and correlated with the clinical severity of blunt emotions. Tron et al. extracted the activity level of 23 facial AUs and found the correlation between negative symptoms severity (e.g., blunted affect) and the positive emotion

expression of patients with schizophrenia [37, 184]. Using these facial expressions, the recognition accuracy of schizophrenia reached up to 85%.

In [4] and [185], Bishay et al. proposed a neural network architecture based on a CNN and a deep neural network (DNN), respectively. The proposed architecture is shown in Figure 2.8. CNN models were implemented first to identify the low-level facial features, where the convolution layer takes into account the value of a pixel, as well as its surrounding pixels. Next, the Gaussian Mixture Model (GMM) and Fisher Vectors (FV) were used to encode the low-level facial features. Fisher vector with GMM represent an image by using the gradient vector of the likelihood function, where the physical meaning of the gradient vector is to describe the parameter change direction that better adapts to the data. The results demonstrated that the facial expressions of schizophrenia patients are correlated to the expression scale of CAINS (MAE=2.67) and the negative symptom scale of PANSS (MAE=3.30). Moreover, another study reported the automated measurement of facial expression (positive, negative, natural face) is statistically correlated with blunted affect and alogia for schizophrenia and schizoaffective disorder [55].

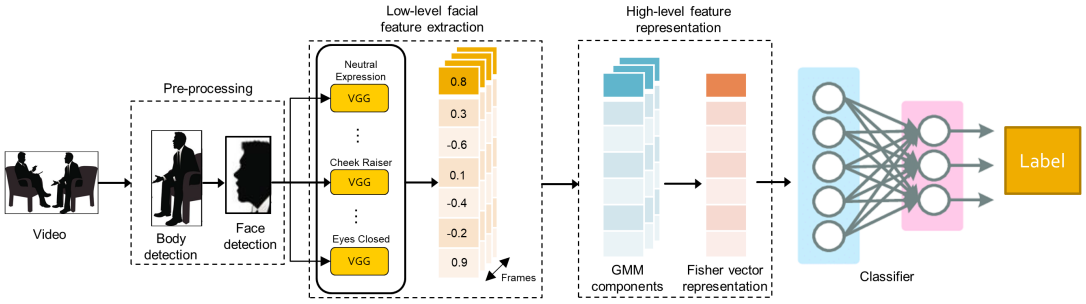


FIGURE 2.8: A CNN-based depression detection pipeline using facial cues.[4]

Visual indicators have been explored for depression detection as well. Most studies attempted to map the facial expression to the severity of depressive symptoms. Based on the FACS AUs, Cohn et al. utilized the manual FACS coding and active appearance modeling to distinguish between depressed and non-depressed participants, and obtained an accuracy of 88% and 79%, respectively [186]. Girard et al. observed that the depression patients with more severe symptoms made fewer affiliative facial expressions and more non-affiliative facial expressions [52]. Similarly, in [187], they found that the smile intensity of distressed patients was lower than non-distressed patients. Girard and Cohn provided a brief overview of audiovisual behavioral analysis of depression [188], where they summarized the

abnormal facial expression of patients with depression including shorter duration of blinks, slower head movements, less head motion, longer duration of looking down, and increased mouth dimpling. The classification accuracy of depression detection generally falls between 70% and 80%, while a few studies reported as high as 90% [188]. The AVEC 2016 [161] scored a mean absolute error of 5.66 for depression severity estimation. In the study by He et al. [51], the authors presented a framework using dynamic facial appearance descriptor and Dirichlet Process Fisher Encoding (DPFE) to evaluate the severity of depressive symptoms. Dibeklioglu et al. proposed a multi-modal approach to predict the level of depression severity using facial and head movements [189]. Meanwhile, Alghowinem et al. constructed low-level and statistical features of head pose and movement for 30 depressed patients and 30 non-depressed patients from facial videos [190]. Through these automated measurements, they achieved an average accuracy of 71.2% in detecting depression. Another review outlines the method and algorithm of feature extraction, dimension reduction, classification/regression methods, and different fusion approaches for automatic depression assessment [50]. It shows that most studies utilized the features extracted from full face, AUs, facial landmarks, and mouth/eyes. The most widely used classification and regression methods are SVM and SVR, respectively. The accuracy of categorical assessment of depression on four data sets (DAIC, BlackDog, Pitt, and AVEC) using facial expression features generally falls between 70% to 100%. Recently, researchers have been developing more complicated deep neural networks to improve the performance of depression detection [191–195]. However, most models are black boxes, and little is known about the patterns captured by the algorithm and which variables are combined to make predictions.

2.4 Body Movement Analysis

Apart from speech deficits and facial expression abnormalities, psychomotor retardation is one core measurement of the negative symptoms [20]. Although many clinical scales are designed to explore the relationship between motor dysfunction and schizophrenia, the development of a method to quantify the motor deficits may contribute to new insights into psychiatric patients [196]. In [41], Kupper and his colleagues leveraged Motion Energy Analysis (MEA) to evaluate the body

movement during the role-play interaction of patients with schizophrenia objectively. The patients who present less body movement are found to be correlated with both negative symptoms and positive symptoms assessed by PANSS. However, role-playing experiments may enhance the participants' physical movement, which is different from the patient's response to normal conversation. MEA was also employed by Ramseyer et al. [197, 198] in order to analyze the synchronicity between the movement of outpatients and therapists. In another study [199], the author collected videos of 31 schizophrenic patients and 32 healthy controls in the Test of Upper Limb Apraxia (TULIA). Through MEA, they found that there are significant differences in quantitative gesture performance between patients and the control group, where patients with schizophrenia need more movement and more time to complete tasks. Moreover, Lavelle et al. measured the nodding and gesturing of a patient with schizophrenia through a 3-D motion capture technology [200]. They found that patients with more severe negative symptoms nod less as listeners and their interlocutors seem to make more gestures as compensation. Walther et al. reported that negative symptoms scores were associated with low activity levels [40]. In addition to body movement and gestures, researchers also investigated that schizophrenia patients often perform poorer body language reading abilities than healthy controls [201] and schizophrenia patients are associated with diminished head motion [52].

Apart from the motor analysis of patients with schizophrenia, a few studies focus on delineating body movement markers for depression. For instance, the upper body and facial movement of patients with depression when responding to the interview questions were analyzed in [54]. They utilized Space-Time Interest Points (STIP) features and achieved the best accuracy of 76.7% when classifying depression and healthy controls. Horigome et al. [53] employed a Red-Green-Blue-Depth (RGB-D) sensor during a clinical interview setting, computed the position, speed, acceleration, and jerk of four body joints, and predicted the severity of depressive symptoms by using machine learning. Another research found that patients with depression presented less movement synchrony and less synchronous positive facial expressions [202]. The MEA evaluates the overall movement of the body, but it requires a static camera position, stable light conditions, and digitized film material. The skeleton-based method is able to measure every part of the body, however, it requires additional equipment, such as a Kinect, to record depth-related

signals. The analysis pipeline of body movement in a clinical interview is shown in Figure 2.9.

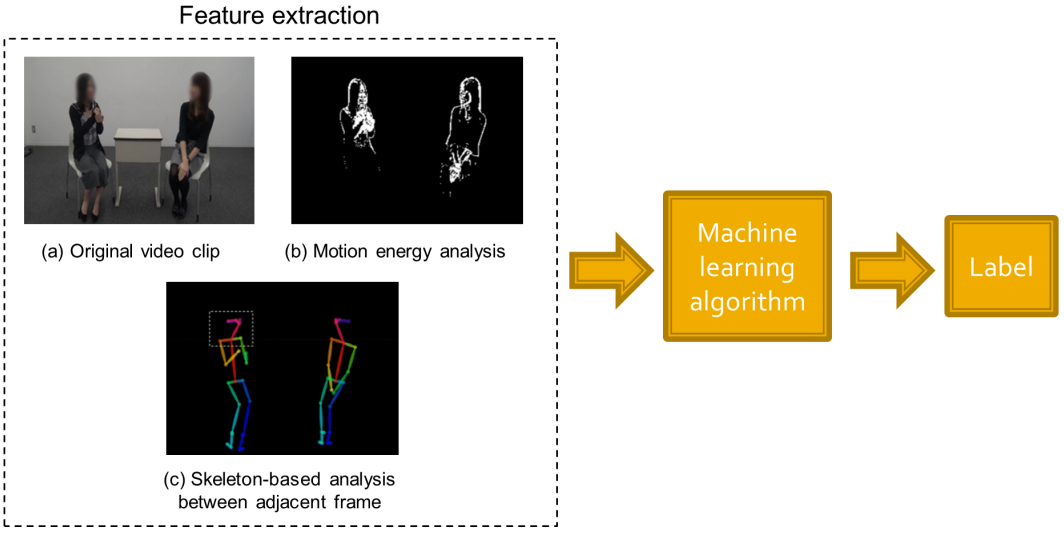


FIGURE 2.9: The analysis methods of body movement in clinical interviews.

However, to the best of our knowledge, except for our preliminary work [39], there is no study leveraging objective measurements of body movement to predict the severity of negative, cognitive, and psychiatric symptoms of schizophrenia and depression patients and differentiate individuals with schizophrenia from healthy controls.

2.5 Speech Adaptation Analysis

Although speech adaptation exists in human-human interaction for decades, it is challenging for humans to evaluate this subtle phenomenon. Several signal-derived approaches have been raised in the literature to model this phenomenon in the recent 10 years. The main processing method of speech adaptation is the turn-taking method, defined as the speech adaptation patterns extracted at the level of the inter-pausal unit (IPU) [203]. For example, Weise and co-workers [204] measured adaptation of pitch, intensity, and speech rate from adjacent speech units. In [91], the authors investigated turn-by-turn f0 accommodation and explored the prosodic parameters locally in the overlapped turn. In addition to forming the speech adaptation at the turn level, Brian Vaughan and his colleague proposed

a Time Aligned Moving Average (TAMA) method to average non-verbal features over a set of overlapping windows [205]. Using the TAMA method, researchers investigated prosodic accommodation in 3 informal conversations[206], 41 Japanese dyadic telephone conversations [207], 6 interviews of psychiatrist and depression patients [208], and 16 audio recordings between client and therapist [209]. We demonstrate the turn-taking-based and TAMA adaptation methods in Figure 2.10. The turn-taking-based measurement is designed to quantify the speech interaction in a natural way, while the TAMA-based method takes the continuous behavior into account.

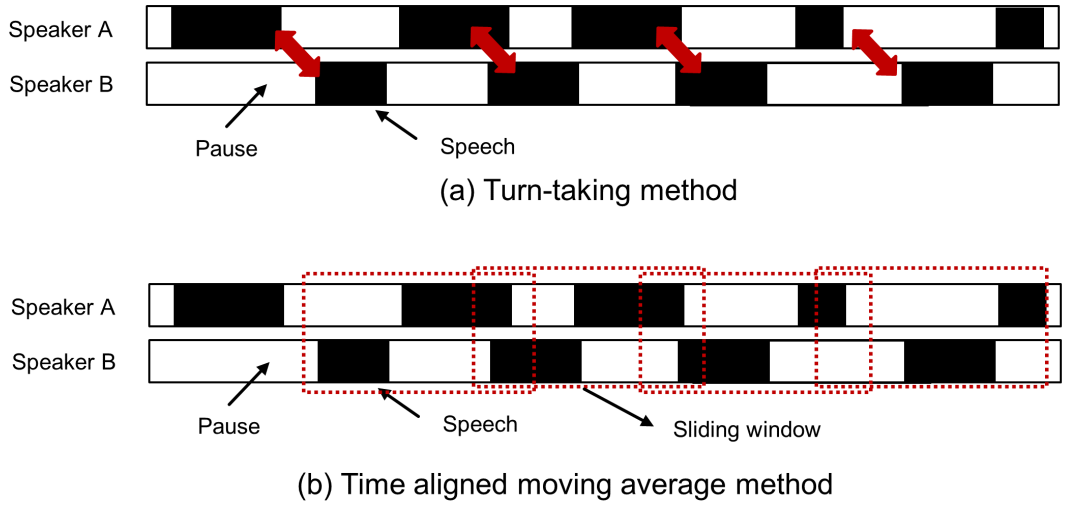


FIGURE 2.10: Illustration of turn-based and TAMA speech adaptation methods.

Moreover, the researchers tried to understand the speech adaptation by learning from real and fake speech turns, where the fake speech turns are usually made up of the same speaker from non-adjacent speech segments. In [207], the authors found significant differences in speech adaptation between real and fake conversations using z-score transformation analysis. Chi-Chun Lee et al. further utilized the eigenvalue of principal component analysis (PCA) to generate the quantitative descriptors of the degree of speech adaptation. They also verified that adaptation is significantly higher in real conversations than in fake conversations [210]. Similarly, linear discriminant analysis (LDA) was also applied to quantify speech adaptation by maximizing the difference between real and fake conversations on a turn-by-turn basis [211]. Furthermore, a DNN-based representation learning method was also proposed to model speech adaptation [5]. Finally, this study embedded the original

acoustic features of one utterance into a low-dimensional embedding, use the differences of two embeddings extracted from two utterances to quantify the speech adaptation, and achieved state-of-art results in separating real and fake conversations. The proposed DNN-based encoder is able to capture the nonlinearity of the speech adaptation and represent the relevant information to a high dimensional space, as shown in Figure 2.11.

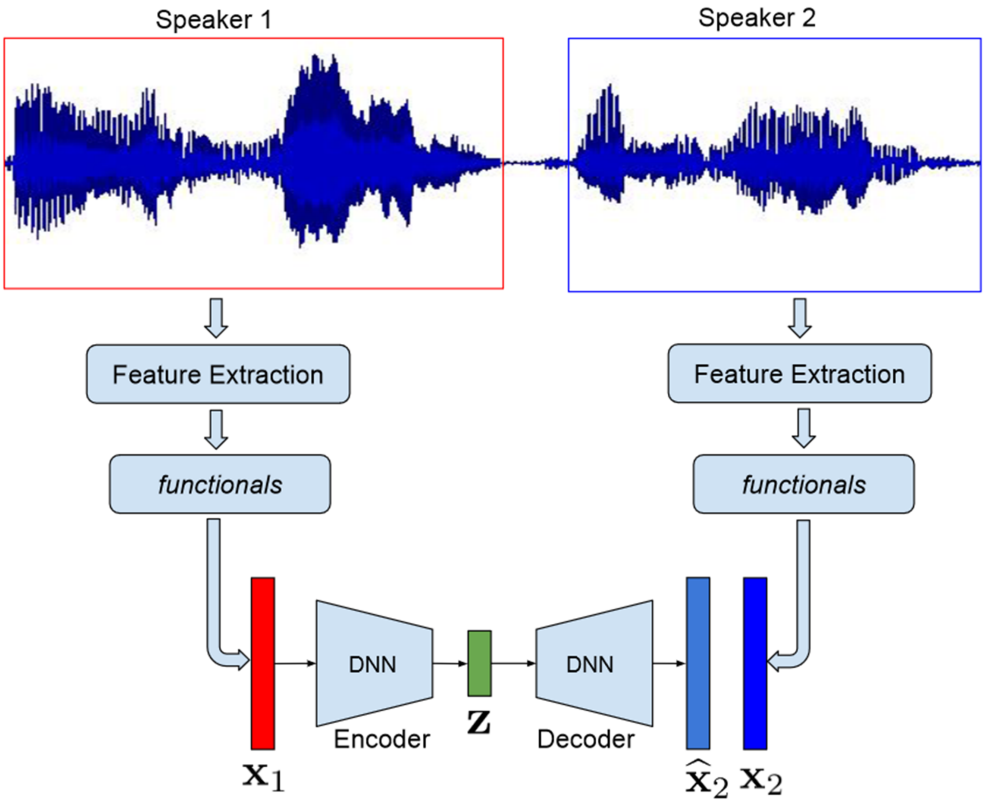


FIGURE 2.11: Illustration of the DNN-based encoder for encoding possible interactive information from speaker turn [5].

In early efforts in understanding speech adaptation, researchers found that it is often related to human mental activities, such as empathy of clinical providers [212] and interpersonal agreement in couples therapy [213]. However, few researchers have addressed how speech adaptation is present in mental disorders and whether the speech adaptation patterns can be used to diagnose or assess patients. One prior work attempted to find prosodic accommodation in clinical interactions between depression patients and psychiatrists [208], but this method is limited in scope because the sample size used in this study is small and did not compare with healthy controls. Therefore, in our work, the pilot data from the recording of both

psychiatrists and patients were utilized to analyze the association between their speech behaviors.

There is an increasing need for methods that allow monitoring beyond health-care settings across time in psychiatry. Voice abnormality in schizophrenia and depression has been wildly researched in the past decade [32]. However, few studies investigate interpersonal behaviors and dynamics between the patients and interlocutors. It is not yet known how adapted vocal patterns could be used to diagnose and monitor mental disorders such as schizophrenia and depression.

2.6 Summary

As described in previous sections, through audio and video analysis, numerous studies have demonstrated that patients suffering from schizophrenia or depression exhibit abnormalities in speech, language, and motor behaviors compared with healthy controls [32, 49, 121]. These abnormal behaviors or digital biomarkers often provide insights into diagnosis and symptom evaluation for psychologists. In summary, the candidate presents the studies used in schizophrenia and depression identification in Table 2.1, the significant digital biomarkers for identifying patients from healthy controls in Table 2.2, and the benefits and limitations of the data-driven methods in Table 2.3. However, our understanding of the relationship between those digital behavioral cues and the mental illness patients and their symptoms is still limited. Whether it is possible to distinguish between different patient groups and the severity of symptoms automatically still requires more studies [110]. Therefore, our team collected the real patient data, and the candidate designed an automatic pipeline to analyze the multiple kinds of behavioral cues, multiple types of symptoms, and multiple types of mental disorders simultaneously. The candidate follows the state-of-the-art methods to build automated data-driven pipelines for diagnostic group classification and symptom severity prediction. The candidate also tries to find out what are the abnormal digital biomarkers for identifying patients to gain insights into the diagnosis and symptoms evaluation of psychologists.

TABLE 2.1: Tabulation of the studies used in schizophrenia and depression identification.

Task	Method	Modality	Accuracy/Reference		
			Below 80%	80% to 90%	Above 90%
SCZ vs. HC	Almost all studies utilized traditional classifiers (e.g., SVM, Logistic Regression, and Random Forest)	Verbal (e.g., Lexical, LSA-based, semantic, graph-based, word2vec-based, and dictionary-based features)	[115], [214], [215] [126], [30]	[127], [117], [121], [216], [128]	[217]
		Non-verbal (e.g., acoustic, prosodic, and conversational features)	[34], [142]	[58], [218]	[143], [219]
		Facial expression (e.g., facial action units and facial cluster features)	-	[184], [37]	-
		Body movement (e.g., joint speed and acceleration)	-	[39]	-
MDD vs. HC	Traditional classifiers (e.g., SVM and Logistic Regression) and deep learning methods (e.g., CNN and LSTM)	Verbal (e.g., sentence embedding)	[141]	[220]	-
		Non-verbal (e.g., acoustic, prosodic, and conversational features)	[221], [222], [223], [224], [161]	[173], [225], [226], [227], [163], [169]	[157]
		Facial expression (e.g., facial action units and head movement features)	[189], [190], [186], [163], [161], [62], [54]	[220], [224], [169], [166]	-
		Body movement (e.g., head motion, joint speed, and joint acceleration)	[53], [54]	-	-

Abbreviation: SCZ=Patient with schizophrenia, HC=Healthy control, MDD=Patient with major depressive disorder, LDA=Linear Discriminant Analysis, LSA=Latent Semantic Analysis, SVM=Support Vector Machines

TABLE 2.2: Significant digital biomarkers observed in the literature.

Modality	Schizophrenia vs. healthy controls	Depression vs. healthy controls
Verbal [44, 45, 115, 126, 135, 137, 228–230]	Number of words per sentence Social-driven words First-person singular pronouns WH-family (e.g., which and what) Adjectives	First-person singular pronouns Negatively emotional words
Non-Verbal [35, 49, 55, 58, 142, 148, 149, 231–233]	Total time talking Speech rate Pause duration Pitch variability Jitter	Pitch mean and range Pitch variability Jitter and shimmer Formant Spectral features (e.g.,MFCC and LSF)
Facial expressions [55, 176, 186, 188, 234]	Proportion of positive expressions Facial action units	Smile intensity and duration Facial action units Blinks Head movements Mouth and lips
Body movement [198, 199, 202]	Less body movement	Less movement synchrony

Compared to the methods utilized in the literature, in this thesis, the candidate implements multiple commonly-used tools and methods to extract the behavioral features, including verbal analysis (LIWC, Diction, topic model, and document embedding), non-verbal analysis (OpenSmile and DisVoice), and facial expression (Affectiva, OpenFace, and Opsis). This study of multi-modal behavior cues of schizophrenia complements the research in the field and is one of the few to automatically analyze the facial expressions of people with schizophrenia. The candidate also expands a few of his seniors' works on a larger corpus, such as measuring conversational cues and 3D body movement. In addition, in this thesis, the authors propose a new quantitative method to assess patients' and psychiatrists' nonverbal interactions during interviews, see chapter 6. The proposed method measures three forms of speech adaptation (matching, reciprocity, and convergence), which are common patterns in human-human interaction but have not yet been used to analyze mentally-ill people. For the machine learning algorithm, the traditional machine learning algorithms (e.g., SVM, logistic regression, random forest, etc.) are the most commonly used algorithms in the literature. Compared to the most recent deep learning methods, the traditional machine learning algorithms have fewer parameters, good performance on the small datasets (in our corpus, there are less than 100 samples in each class), and have higher interpretability. Therefore, in this thesis, an ensemble classifier, which consists of multiple traditional machine learning algorithms, is designed to integrate multiple feature sets to automatically classify the diagnostic groups and predict the symptom severity for patients with schizophrenia and depression. Finally, the candidate also found a number of significant digital biomarkers that are helpful in distinguishing different diagnostic groups. Some biomarkers are consistent with previous studies, while some of them are unique findings. Please refer to Section 5.5 for details.

TABLE 2.3: Strengths and limitations of methods for automated diagnosis or assessment of schizophrenia and depression.

Domain	Method/Algorithm	Benefits/Characteristics	Limitations
Verbal	Bag-of-words methods (e.g., td-idf)	• Able to quantify the importance or relevance of string representations (words, phrases, lemmas, etc.) in a document	• It is slow when the vocabulary is large • It makes no use of semantic similarities between words
	Human defined dictionary-based methods (e.g., LIWC and Diction)	• Pre-defined by psychologist • Meaningful categories related to the thoughts, drive, and emotion	• Did not capture the different semantic meanings of words • Human-defined categories are limited
	Topic modeling (e.g., LSA and LDA)	• Foundational techniques in topic modeling • Automatically extract topics and the probability of the topics	• Lack of interpretable embeddings (we don't know what the topics are) • Require a large set of documents and vocabulary to get accurate results • LSA assumes a Gaussian distribution of the terms in the documents, which may not be true for all problems
	Document embedding (Doc2Vec)	• It takes word order into account, generalizes to longer documents, and can learn from unlabelled data • Works for long documents	• Hard to get generally meaningful embeddings if the corpus is small • The meaning of embeddings is not explicitly defined, and there is litter theoretical support behind such characteristics
	Word embedding (e.g., word2vec, GloVe, ELMo, FastText, and BERT)	• Generate word representations from the large corpus • BERT includes attention mechanisms to detect context between words	• Hard to train when the number of categories is too large • Lack of interpretable embeddings (high dimensional representation) • Hard to get generally meaningful embeddings if the corpus is small
Non-Verbal	Low-level-descriptors (e.g., pitch, intensity, formant, and spectral features) extracted from tools (e.g., OpenSmile, Covarep, and Praat)	• Validated on many studies and corpus • Have physical meaning for each feature	• The features generated using statistical functions (e.g., min, max, mean, std) often contain redundant information
	Deep-learning-based methods (e.g., CNN) extracted vocal features from Mel-spectrogram	• Mel-spectrogram is a way to represent a signal's loudness at different frequencies visually • CNN can capture the smaller and simpler patterns through the convolution layer and select the important features through the pooling layer • Often higher performance than hand-crafted features in many audio classification tasks	• Hyper-parameter tuning • Lack of interpretability • Can only capture features in Mel-spectrogram
Facial Expression	Hand-crafted features (e.g., emotions, facial action units, eye-gaze, and head movement) extracted from tools (e.g., Affectiva and OpenFace)	• Interpretable • Easy to find and filter out the interferences	• The features generated using statistical functions (e.g., min, max, mean, std) often contain redundant information • Require the prior knowledge
	End-to-end deep-learning-based methods (e.g., CNN and DNN)	• CNN can capture the smaller and simpler patterns through the convolution layer and select the important features through the pooling layer • Higher performance than hand-crafted features	• Required large corpus • Easy to overfit on external data sets
Body Movement	MEA – motion energy analysis	• An objective automated method that continuously monitors the amount of movement occurring in pre-defined regions of interest.	• The basic prerequisites for MEA are a static camera position, stable light conditions, and digitized film material
	Computed the position, speed, and acceleration of the body skeleton data	• Enabled to capture the 3D movement of the body	• Requires additional equipment, such as a Kinect, to record depth-related signals
Speech Adaptation	IPU-based methods (Inter-Pausal Unit)	• Measured the speech interaction between the turn-takings	• Have to define detect the speaking turns
	TAMA-based methods (Time Aligned Moving Average)	• A continuous method for measuring trends in two interlocutors	• Difficult to select the time window • Affect by the pause in the conversation
	DNN-based methods that evaluate the real and fake conversation	• Captured the non-verbal interaction features in the real human conversation	• Lack of interpretability of what type of speech adaptation features were captured

Chapter 3

Experimental Design

This Chapter first introduces the participants recruitment, demography, and the experimental setup. Second, the subjective assessment scales and group-level differences are presented to provide more clinical insights. This chapter then presents how to binarize the clinical assessment scores and explains the proposed ensemble learning model that integrates those numerous features. Lastly, this chapter introduces the evaluation metrics and explains how the importance of various features between different groups of participants was calculated.

3.1 Participants

This study is cooperated with the Institute of Mental Healthy (IMH) in Singapore. Our team recruited 103 patients with schizophrenia, 50 patients with major depressive disorder, and 75 healthy controls in two studies in Singapore (flow diagram in Figure 3.1). A subset of 54 patients with schizophrenia and 26 health controls were recruited and assessed in the first study (Study-A) between 2014 to 2015 over three sessions to verify the effectiveness of Cognitive Remediation Therapy [235] initially at week 0, week 2, and week 12 (Study-A1, Study-A2, Study-A3). Another subset of 49 patients with schizophrenia, 50 patients with depression, and 49 healthy controls were recruited in the second study (Study-B) under the same protocol between 2017 and 2018. Data collected in week 0 (schizophrenia=54, healthy=26) are merged with the data collected in the second study (schizophrenia=49, depression=50, healthy=49) as one entire data set to train our machine

learning algorithm (demographics in Table Table 3.1), whereas the data collected in week 2 and week 12 were considered as validation sets (Figure 3.1).

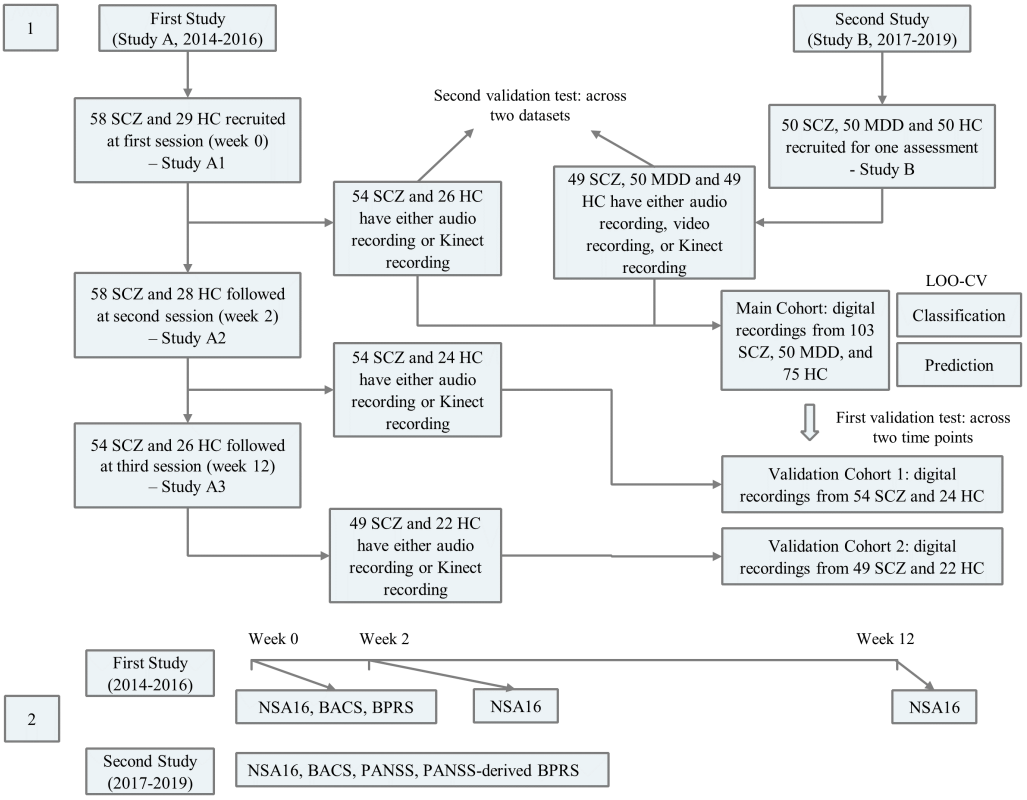


FIGURE 3.1: Patients flow diagram and assessments of two studies.

In both studies, patients were from the Institute of Mental Health Singapore (IMH), and healthy participants were recruited from the general population. The schizophrenic patient groups in both studies were matched for the severity of symptoms (Table 3.2). In addition, all groups of participants are matched for age, gender, educational background, and ethnicity. Patients with schizophrenia were prospectively selected for persistent and predominantly negative symptoms with minimal positive symptoms. For patients with depression, only patients with mild depressive symptoms were chosen to participate in the experiment. Participants were excluded if they had a history of strokes, traumatic brain injuries, neurological disorders like epilepsy, or autistic spectrum disorder. Besides, the participants who attended the first study were excluded from the second study. All patients were evaluated based on the Structured Clinical Interview (SCID) of the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) by well-trained psychologists [236]. Ethical approval was obtained from the field-specific

TABLE 3.1: Descriptive analysis of schizophrenia and healthy group in the first and second study.

	Depression (N = 50)	Schizophrenia (N = 49)	Healthy Controls (N = 49)	Tukey's HSD test		
				P_{SH}	P_{DH}	P_{DS}
Age (years)	36.66 ± 12.88	35.65 ± 10.09	36.37 ± 12.04	0.900	0.900	0.853
Gender (male:female)	26:24	50:53	38:37	0.900	0.900	0.900
Ethnicity(Chinese:Malay:India:Others)	36:5:6:3	87:7:9:0	54:16:4:1	0.514	0.427	0.059
Education years	14.20 ± 2.57	13.58 ± 2.79	13.54 ± 2.66	0.900	0.381	0.388
Duration of illness (years)	4.80 ± 4.89	11.63 ± 9.17	-	-	-	<0.005
Medication						
CPZ equivalence (mg/day)	-	504.18 ± 410.94	-	-	-	-
AntiDDosage (mg/day)	77.00 ± 65.32	-	-	-	-	-
NSA Total Score	40.38 ± 7.71	41.67 ± 9.55	30.12 ± 6.53	<0.005	<0.005	0.630
NSA-Restricted Speech	3.22 ± 1.40	3.37 ± 1.78	2.61 ± 1.23	<0.05	0.080	0.820
NSA-Poor Quality of Speech	2.91 ± 1.08	3.25 ± 1.27	2.13 ± 0.91	<0.005	<0.005	0.186
NSA-Affective Blunting	6.24 ± 2.23	6.23 ± 2.62	4.02 ± 1.53	<0.005	<0.005	0.900
NSA-Amotivation	9.59 ± 2.02	9.29 ± 2.23	5.60 ± 1.91	<0.005	<0.005	0.668
BACS Composite Score	-0.16 ± 1.22	-1.74 ± 1.42	0.05 ± 1.18	<0.005	<0.005	<0.05
BPRS Total Score	32.72 ± 5.58	32.85 ± 8.26	20.39 ± 2.01	<0.005	<0.005	0.900
BPRS-Affective	12.78 ± 3.44	8.53 ± 3.13	5.64 ± 1.33	<0.005	<0.005	0.010
BPRS-Positive	4.51 ± 1.12	7.79 ± 3.89	3.77 ± 0.18	<0.005	0.500	<0.005
BPRS-Negative	7.50 ± 2.06	7.55 ± 2.69	4.78 ± 0.86	<0.005	<0.05	0.056
BPRS-Resistance	4.32 ± 1.38	5.04 ± 1.83	3.37 ± 0.54	<0.005	0.662	<0.005
PANSS Total Score	51.36 ± 8.41	55.61 ± 12.57	35.10 ± 3.75	<0.005	<0.005	0.055
PANSS-FSNS	12.96 ± 5.73	13.71 ± 6.59	8.47 ± 6.15	<0.005	<0.05	0.219
PANSS-DE	4.29 ± 1.53	4.85 ± 2.11	3.25 ± 1.06	<0.005	<0.005	0.900
PANSS-SA	5.12 ± 1.66	5.00 ± 1.56	2.96 ± 0.84	<0.005	<0.005	0.522
Number of Recordings						
Audio	48	98	70	-	-	-
Video	42	44	45	-	-	-
Kinect	42	92	66	-	-	-
Audio or Video	50	99	74	-	-	-
Audio or Kinect	50	103	75	-	-	-
Video or Kinect	42	92	66	-	-	-
Audio or Video or Kinect	50	103	75	-	-	-

Values are shown as mean ± SD. The group demographics and assessment scores are compared by t-tests. Abbreviation: P_{SH1}=p-value between schizophrenia and control groups of the first study; P_{SH2}=p-value between schizophrenia and control groups of the second study; P_{SS}=p-value between schizophrenia groups of two studies; CPZ=Chlorpromazine; BACS=Brief Assessment of Cognition in Schizophrenia; BPRS=Brief Psychiatric Rating Scale-18; NSA=16-item Negative Symptoms Assessment; mg=milligram; NA=Not Applicable.

review committee of the National Healthcare Group Singapore. All participants were briefed on their participation and withdrawal rights, provided written informed consent, and received monetary reimbursement for their participation.

3.2 Experimental Setup

The experimental setup is illustrated in Figure 3.2. Specifically, audio and Kinect skeleton data were recorded during the semi-structured interview of both Study-A and Study-B. The voice of both the participant and the psychiatrist were recorded through two separate lapel microphones. These two microphones were connected to an H4N recorder which captures the two-channel speech signals at 48kHz. The RGB and depth data were also recorded through Microsoft Kinect v1 and v2 for

TABLE 3.2: Descriptive analysis of schizophrenia and healthy group in the first and second study.

	First Study		Second Study		P_{SH1}	P_{SH2}	P_{SS}
	Schizophrenia (N = 54)	**Controls (N = 26)**	**Schizophrenia (N = 49)**	**Controls (N = 49)**			
Age (years)	31.22 ± 7.52	29.58 ± 7.93	40.53 ± 10.31	39.98 ± 12.28	0.377	0.812	<0.005
Gender (male:female)	25:29	12:14	24:25	23:26	0.991	0.842	0.788
Education years	13.65 ± 2.75	13.53 ± 2.18	13.50 ± 2.83	13.54 ± 2.88	0.849	0.948	0.794
Duration of illness (years)	9.06 ± 7.30	NA	14.47 ± 10.14	NA	NA	NA	<0.005
CPZ equivalence (mg/day)	442.97 ± 345.72	NA	567.36 ± 466.33	NA	NA	NA	0.139
NSA Total Score	41.39 ± 9.34	26.77 ± 3.69	41.04 ± 10.47	31.90 ± 7.00	<0.005	<0.005	0.860
NSA-Restricted Speech	3.42 ± 1.77	2.15 ± 0.50	3.38 ± 1.97	2.86 ± 1.41	<0.005	0.138	0.917
NSA-Poor Quality of Speech	3.18 ± 1.32	1.59 ± 0.30	3.94 ± 4.22	2.41 ± 0.99	<0.005	<0.05	0.219
NSA-Affective Blunting	5.91 ± 2.44	3.41 ± 0.72	7.14 ± 5.34	4.34 ± 1.74	<0.005	<0.005	0.134
NSA-Amotivation	9.29 ± 2.35	5.13 ± 1.57	9.05 ± 2.54	5.86 ± 2.02	<0.005	<0.005	0.617
BACS Composite Score	-1.86 ± 1.27	0.50 ± 0.99	-1.49 ± 1.75	-0.18 ± 1.21	<0.005	<0.005	0.230
BPRS Total Score	32.76 ± 8.77	19.81 ± 1.82	32.55 ± 8.21	20.69 ± 2.03	<0.005	<0.005	0.904
BPRS-Affective	8.83 ± 3.59	5.28 ± 1.09	8.14 ± 2.56	5.83 ± 1.40	<0.005	<0.005	0.271
BPRS-Positive	7.89 ± 4.06	3.73 ± 0.00	7.64 ± 3.73	3.79 ± 0.22	<0.005	<0.005	0.747
BPRS-Negative	6.87 ± 2.57	4.49 ± 0.64	8.09 ± 2.75	4.94 ± 0.92	<0.005	<0.005	<0.05
BPRS-Resistance	5.26 ± 2.16	3.54 ± 0.72	4.75 ± 1.42	3.27 ± 0.38	<0.005	<0.005	0.170
Number of Recordings							
Audio	50	25	48	45	NA	NA	NA
Video	NA	NA	44	45	NA	NA	NA
Kinect	47	21	45	45	NA	NA	NA
Audio or Video	50	25	49	49	NA	NA	NA
Audio or Kinect	54	26	49	49	NA	NA	NA
Video or Kinect	47	21	45	45	NA	NA	NA
Audio or Video or Kinect	54	26	49	49	NA	NA	NA

Values are shown as mean ± SD. The group demographics and assessment scores are compared by t-tests. Abbreviation: P_{SH1}=p-value between schizophrenia and control groups of the first study; P_{SH2}=p-value between schizophrenia and control groups of the second study; P_{SS}=p-value between schizophrenia groups of two studies; CPZ=Chlorpromazine; BACS=Brief Assessment of Cognition in Schizophrenia; BPRS=Brief Psychiatric Rating Scale-18; NSA=16-item Negative Symptoms Assessment; mg=milligram; NA=Not Applicable.

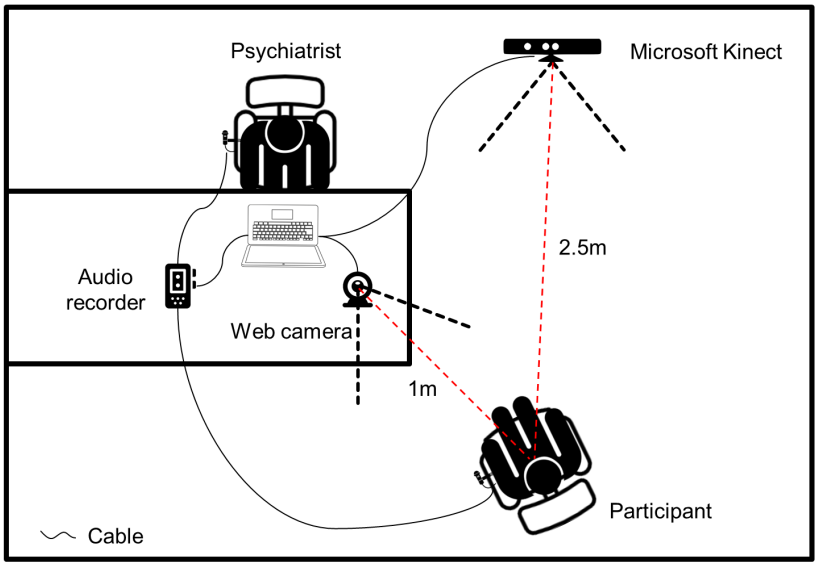
Study-A and Study-B, respectively, while the participant was seated in a fixed position.

Meanwhile, a webcam was directed to the participant's face and recorded video at 1080p resolution (only for Study B). The psychiatrist and the Kinect device's location are approximately 2.5 meters away from the participant, while the webcam was about 1.5 meters away. All the digital recordings were recorded and stored on a laptop by an assistant during the interview.

3.3 Clinical Assessments

All participants have been assessed for negative symptoms through the 16-item Negative Symptoms Assessment (NSA-16) [20] and cognitive deficits through the Brief Assessment of Cognition in Schizophrenia (BACS) [237]. The participants of the first and second study were assessed for general psychiatric symptoms utilizing the Brief Psychiatric Rating Scale-18 (BPRS) [119] and Negative Syndrome Scale (PANSS) [120], respectively. For consistency, for the second study, BPRS scores were derived from the PANSS scores, i.e., the 18 BPRS items included in the

(a) Frontal view.



(b) Top-down view.

FIGURE 3.2: Illustration of the experimental setup.

PANSS were used, which we analyze together with the BPRS scores from the first study. It is important to note that the NSA-16 was assessed during the semi-structured clinical interview, while the other assessments, e.g., PANSS, BPRS, and BACS, were evaluated right after the NSA semi-structured interview.

To facilitate clinical interpretation, the factor scores from the NSA and BPRS scores were also analyzed in this thesis. Specifically, NSA-16 includes four domains: restricted speech (NSA-RS), poor quality of speech (NSA-PQ), affective blunting (NSA-AB), and amotivation (NSA-AM) [238]. Similarly, we consider the four-factor scales of BPRS, including Affective (BPRS-AFF), Positive (BPRS-POS), Negative (BPRS-NEG), and Resistance (BPRS-RES) [239]. In addition, the BACS

subscales and composite scores were standardized by Z-scores according to the scores of healthy controls.

TABLE 3.3: List of factor, domain, total, and individual scales of NSA-16, BACS, BPRS, and PANSS.

Scale	Factor/Domain/Total scale	Individual scale
NSA-16	NSA-RS: Restricted speech	NSA2: Restricted speech quantity
		NSA9: Poor rapport with interviewer
	NSA-PQ: Poor quality	NSA3: impoverished speech content
	of speech	NSA4: Inarticulate speech
	NSA-AB: Affective blunting	NSA1: Prolonged time to respond
		NSA6: Reduced modulation of intensity
		NSA15: Reduced expressive gestures
		NSA16: Slowed movements
	NSA-AM: Amotivation	NSA8: Reduced social drive
		NSA12: Reduced sense of purpose
		NSA13: Reduced interests
		NSA14: Reduced daily activity
	NSA-Total	Sum of all NSA-16 items
BACS	BACS-Composite	Composite score of BACS calculated from the
		z-score of individual BACS items:
		BACS-VM: Verbal memory
		BACS-DS: Digit sequencing
		BACS-TMT: Token Motor
		BACS-SF: Symbol Fluency
		BACS-SC: Symbol Coding
		BACS-ToL: Tower of London
BPRS-18	BPRS-Affective	BPRS1: Somatic concern
		BPRS2: Anxiety
		BPRS5: Guilt feelings
		BPRS6: Tension
		BPRS9: Depressive mood
	BPRS-Positive	BPRS4: Conceptual disorganization
		BPRS8: Grandiosity
		BPRS12: Hallucinatory behavior
		BPRS15: Unusual thought content
	BPRS-Negative	BPRS3: Emotional withdrawal
		BPRS13: Motor retardation
		BPRS16: Blunted affect
		BPRS18: Disorientation
		BPRS7: Mannerisms and posturing

TABLE 3.3: List of factor, domain, total, and individual scales of NSA-16, BACS, BPRS, and PANSS.

Scale	Factor/Domain/Total scale	Individual scale
	BPRS-Resistance	BPRS10: Hostility
		BPRS11: Suspiciousness
		BPRS14: Uncooperativeness
		BPRS17: Excitement
	BPRS-Total	Sum of all BPRS-18 items
PANSS-30	PANSS-POS: Positive	PANSSP1: Delusions
		PANSSP6: Suspiciousness/Persecution
		PANSSG16: Active Social Avoidance
	PANSS-NEG: Negative	PANSSN1: Blunted Affect
		PANSSN2: Emotional Withdrawal
		PANSSN3: Poor rapport
		PANSSN4: Passive social withdrawal
		PANSSN6: Lack of spontaneity and flow
		PANSSN7: Stereotyped thinking
		PANSSG5: Mannerisms and Posturing
		PANSSG7: Motor Retardation
		PANSSG10: Disorientation
		PANSSG12: Lack of Judgment and Insight
	PANSS-COG:	PANSSP2: Conceptual Disorganization
	Cognitive/Disorganization	PANSSP5: Grandiosity
		PANSSG9: Unusual Thought Content
		PANSSG11: Poor Attention
		PANSSG13: Disturbance of Volition
PANSS-30	PANSS-DEP:	PANSSG1: Somatic Concern
	Depression/Anxiety	PANSSG2: Anxiety
		PANSSG3: Guilt Feelings
		PANSSG4: Tension
		PANSSG5: Mannerisms and Posturing
		PANSSG6: Depression
		PANSSG7: Motor Retardation
	PANSS-HOS: Hostility	PANSSP4: Excitement
		PANSSP7: Hostility
		PANSSG8: Uncooperativeness
		PANSSG14: Poor Impulse Control
	PANSS-Total	Sum of all PANSS items

3.4 Group-level Differences

A total of 228 patients and healthy controls (HC) were included in the analysis (mean [SD] age, 36.11 [11.44] years; 114 [50%] female and 114 [50%] male) from two studies. This thesis analyzed the audio and video recordings of 103 patients with schizophrenia, 50 patients with depression, and 75 HC matched for age, gender, ethnicity, and educational background (Table 3.1). The average illness duration [SD] of patients with schizophrenia is 11.63 (9.17) years and 4.80 (4.89) years for patients with depression. Compared with the healthy group, both schizophrenia and depression patients had significant differences in negative symptoms (NSA-Total of patients with schizophrenia: mean [SD], 41.67 [9.55], P_{SH} <0.005; patients with depression: mean [SD], 40.38 [7.71], P_{DH} <0.005), cognitive symptoms (BACS-Composite of patients with schizophrenia: mean [SD], -0.16 [1.22], P_{SH} <0.005; patients with depression: mean [SD], -1.74 [1.42], P_{DH} <0.005), and general psychiatric symptoms (BPRS-Total of patients with schizophrenia: mean [SD], 32.85 [8.26], P_{SH} <0.005; patients with depression: mean [SD], 32.72 [5.58], P_{DH} < 0.005). When comparing the schizophrenia and depression groups, there was no statistically significant difference in the severity of overall negative symptoms (NSA-Total, P_{DS}=0.63). The cognitive symptoms of the schizophrenia group were more severe than the depression group (BACS-Composite, P_{DS} <0.05). For general psychiatric symptoms, although the overall BPRS score did not differ statistically between schizophrenia and depression patients (BPRS-Total, P_{DS} =0.90), patients with depression showed more anxiety and depression (BPRS-AFF, P_{DS} <0.005), and the patients with schizophrenia had higher ratings on BPRS-POS (P_{DS} <0.005) and BPRS-RES (P_{DS} <0.005).

3.5 Label Binarization

The classification task defines patients with different diagnosis results as distinct categories, while the prediction task has a more detailed distinction on the symptom severity. Most of the patients in our study have only mild symptoms. Therefore, the symptom scores do not cover the entire range but typically take low values instead. Those scores were divided into two classes to predict the scores, distinguishing the

severity of symptoms on two levels only. In other words, each subjective rating was split into class *Low* (score $<$ threshold) and class *High* (score $\geq$ threshold).

For most of the ratings, the median values on the training data were selected as the cut-off score such that both classes have similar counts. In this manner, the data is well balanced between the two categories. To address the clinical significance, this thesis also evaluated the cut-off scores of PANSS-FSNS, PANSS-Total, and BPRS-Total utilizing the equipercentile linking results [240–242]. For instance, the cut-off scores of PANSS-Total and PANSS-FSNS between normal and borderline illness are set to 38 and 9.5 respectively; the cut-off scores of PANSS-Total and PANSS-FSNS for borderline and mild illness is set to 52 and 14.5 respectively; the cut-off scores of BPRS-Total between normal and borderline and between borderline and mild illness are set to 24 and 32, respectively. For cognitive symptoms, the BACS-composite values of -1 and -2, which represent one and two standard deviations relative to healthy people, were leveraged to determine normal (score $>$ -1), mild (-1 $<$ score $<$ -2), and severe (score $<$ -2) cognitive symptom [243].

3.6 Classification Method

The objective of this thesis is to predict subjective assessment scores (prediction tasks) from those numerous features and to classify the three different participant groups (classification tasks). For the sake of comparison, similar studies that report classification and prediction results using machine learning techniques were summarized in Table B.1 and Table B.2 respectively. In this study, an ensemble learning pipeline was designed for the prediction and classification tasks, implemented in the Scikit-learn toolkit (version 0.23.2) in Python 3.8.[244] We depict the ensemble learning pipeline in Figure 3.3.

All classification and prediction tasks were validated through leave-one-out cross-validation (LOO-CV). In each LOO-CV loop, one participant was held out as the test sample. The other samples made up the training set for training the classifiers. As illustrated in Figure 3.3 (a), a separate ensemble classifier was trained for each feature set separately. Each of those ensemble classifiers contains five *base learners*: Support Vector Machine (SVM) with linear kernels, Logistic Regression, Gradient Boosting, AdaBoost, and Random Forest. The hyperparameters and the random

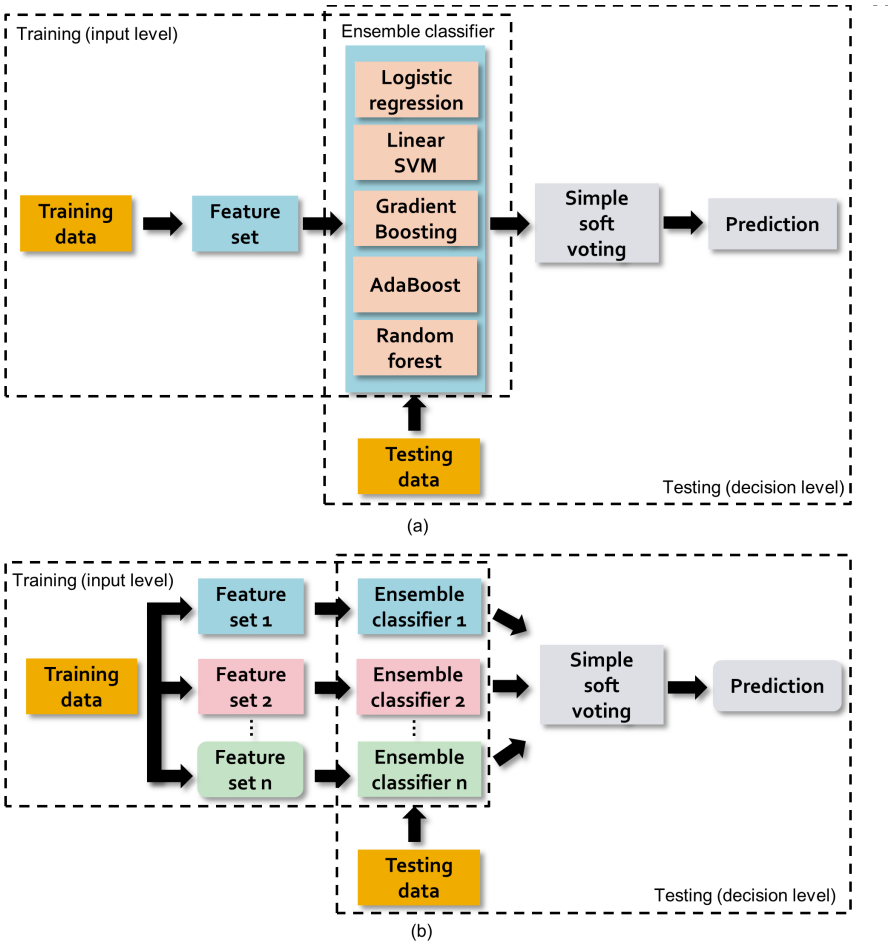


FIGURE 3.3: Ensemble learning pipeline.

seeds were fixed for those five base classifiers in order to generate reproducible results. The hyperparameters are listed in Table 3.4. Next, each base learner makes its own predictions, and the predictions of the 5 base learners are averaged to generate the final prediction for a given feature sets. Before combining the outputs of the base classifiers, we standardized those predictions from each feature set in a non-trivial manner (referred to as probability calibration). An internal LOO-CV was first applied to obtain the probability outputs on the training set. Next, the minimum, maximum, and optimal threshold of these probability outputs on the training set were used to calibrate the predictions of the test set into a range of 0 and 1 for each of the 5 base classifier (Figure 3.5), where the optimal threshold is determined as the decision threshold with the maximum geometric mean score (G-mean). Lastly, as illustrated in Figure 3.3 (b), the standardized predictions from all feature sets are combined by averaging, resulting in the final prediction based on all feature sets.

TABLE 3.4: Key hyperparameters of all 5 base classifiers in the ensemble classifier.

Classifier	Parameter name	Value	Description of parameter
Linear SVM	C	1.0	Regularization parameter
	class_weight	balanced	Weights of each class are automatically adjusted as n_samples/(n_classes*n_samples_in_class)
	max_iter	100000	Hard limit of 100000 iterations
Logistic Regression	solver	L-BFGS	L-BFGS is an optimization algorithm that approximates the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) using a limited amount of computer memory
	penalty	L2	L2 regularization
	C	1.0	Regularization parameter
	class_weight	balanced	Weights of each class are automatically adjusted as n_samples/(n_classes*n_samples_in_class)
	max_iter	100000	Hard limit of 100000 iterations
Gradient Boosting	loss	deviance	Use binomial deviance loss for classification with probabilistic outputs
	learning_rate	0.1	Learning rate shrinks the contribution of each tree
	n_estimators	100	The number of boosting stages to perform
	max_depth	3	Maximum depth of the individual regression estimators
AdaBoost	n_estimators	100	The number of boosting stages to perform
	algorithm	SAMME.R	Real Stagewise Additive Modeling using a Multi-class Exponential (SAMME.R) loss function
	learning_rate	1.0	Learning rate shrinks the contribution of each tree
Random Forest	n_estimators	400	The number of trees in the forest
	criterion	Gini	The Gini impurity is used to measure the quality of a split
	class_weight	balanced	Weights of each class are automatically adjusted as n_samples/(n_classes*n_samples_in_class)
	max_depth	7	The maximum depth of the tree

For some of the classification and prediction tasks (e.g., depression and schizophrenia), there are considerably more samples from one class than the other. To overcome this class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) [245] was applied to create synthetic data for the minority class by interpolating existing data points.

The z-score standardization was then applied to all features, which subtracts the mean from each feature value and divide by the standard deviation. As a result, the standardized features have a mean of 0 and a standard deviation of 1.

In the multi-class classification tasks (schizophrenia vs. depression vs. HC), the one-versus-the-rest approach was applied, where for each class a classifier is trained to distinguish that class from the others (e.g., depression vs. schizophrenia/controls), leading to an output for each class.

3.7 Performance Evaluation

To evaluate the classification and prediction performance, several standard classification metrics were calculated: confusion matrix (CM), sensitivity (SEN), specificity (SPE), accuracy (ACC), balanced accuracy (BAC), weighted F1-score (F1), the Matthews correlation coefficient (MCC), and area under the precision-recall

curve (AUPRC). In addition, the majority baseline (MB) was considered as the benchmark, which assigns the majority class (most frequent category) to all samples. All analyses were performed using Python 3.8 and Scikit-Learn toolkit 0.23.2 [244]. Additionally, the random seed was fixed in Python to obtain reproducible results.

The confusion matrix provides a clear visualization for the performance of a classification result. The binary CM uses four types of metrics, named true positive (TP), false negative (FN), false positive (FP), and true negative (TN), to indicate the classification results. These four metrics can be represented in a 2×2 CM, as shown in Table 3.5. For example, TP represents the number of samples in positive class predicted to be positive.

TABLE 3.5: Example of a confusion matrix for a binary classification task.

Confusion Matrix		Predicted label	
		Positive	Negative
True label	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

SEN, SPE, F1, ACC, BAC, and MCC are commonly used indicators for evaluating classification performance. We summarized their calculation formulas in Table 3.6. The value range of all these parameters is between 0 and 1, and the larger the value, the better the classification performance.

TABLE 3.6: Equations of evaluation metrics.

Metric	Formula
Sensitivity (SEN)	$SEN = \frac{TP}{TP + FN}$
Specificity (SPE)	$SPE = \frac{TN}{TN + FP}$
F1-Score (F1)	$SPE = \frac{TN}{TN + FP}$
Accuracy (ACC)	$F1 = \frac{2 * TP}{2 * TP + FP + FN}$
Balanced Accuracy (BAC)	$BAC = \frac{SEN + SPE}{2}$
Matthews correlation coefficient (MCC)	$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$

The AUPRC metric used in this thesis is illustrated in Figure 3.4. The average value of the area under two blue curves is the AUPRC metric. AUPRC is a better metric for imbalanced classification compared to area under receiver operating characteristic curve (AUROC) [246].

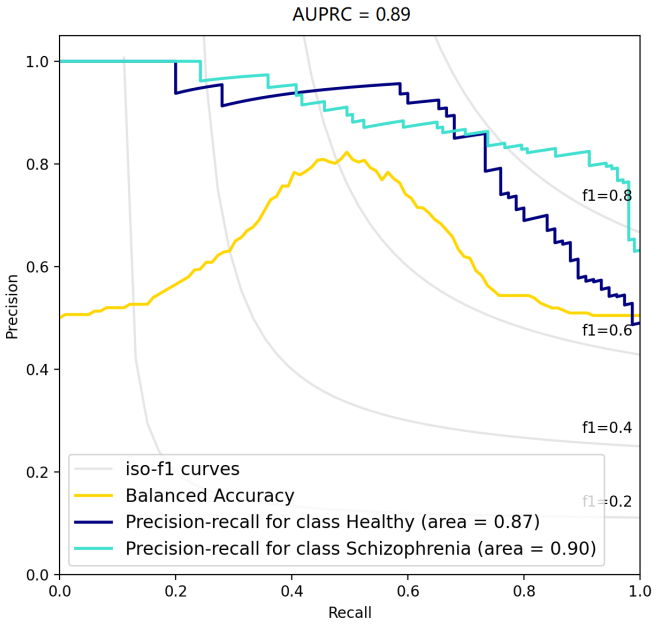


FIGURE 3.4: Illustration of the precision-recall curve and AUPRC in a schizophrenia and health classification task.

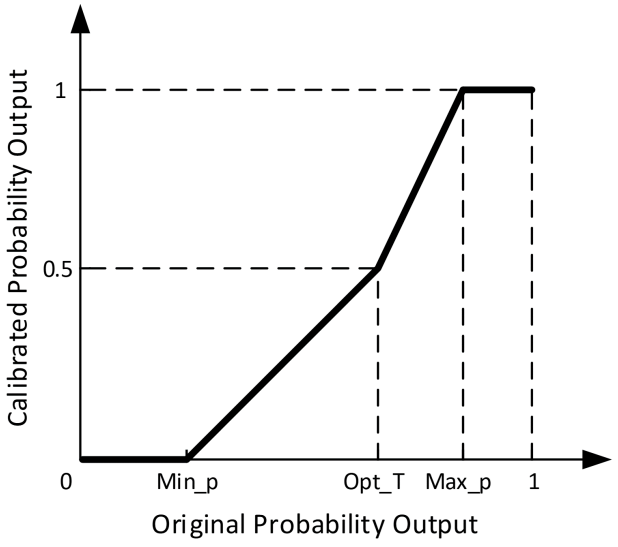


FIGURE 3.5: Probability calibration using a piecewise linear function.

3.8 Feature Importance Measurement

As explained in earlier sections, a plethora of features were extracted from the audio and video recordings to predict clinical scores and classify the participants. To better understand which features are the most predictive, both statistical and model-based methods were applied to rank the features.

Since an ensemble of five base classifiers was trained (Logistic Regression, linear SVM, AdaBoost, Gradient Boosting, and Random Forest) for each feature set, the model-specific importance of each feature for each of the 5 base classifiers was first evaluated employing the Scikit-learn toolkit [244]. Next, the importance coefficients were normalized by min-max scaling and then took the average of the normalized weights; this average normalized weight measures the importance of the feature at hand. Lastly, the features were ranked according to this average normalized weight. For audio and video feature sets, this thesis listed the top 5 features according to this ranking (see Sections 4.6 and 5.5). The top 10 features of speech adaptation features were presented in Section 6.4. For the sake of completeness, the method that obtains the feature weight for the 5 base classifiers was briefly explained in the following. In the Logistic Regression classifier, the absolute value of the regression coefficient represents the importance. In the linear SVM model, the vector perpendicular to the classification hyperplane represents feature importance, while tree-based models (AdaBoost, Gradient Boosting, and Random Forest) measure feature importance as the average reduction in impurity brought by that feature [244, 247].

This thesis also indicated whether those features are significant according to the Kruskal-Wallis test. Since many different modalities and features were utilized simultaneously, leading to multiple statistical tests, the statistical post-correction was applied to the p-values of the Kruskal-Wallis tests. Specifically, the Benjamini-Hochberg False Discovery Rate (FDR) post-correction was applied to all Kruskal-Wallis p-values simultaneously using the MNE toolkit[1] in Python 3.8, resulting in corrected p-values. Based on those corrected p-values, we determine whether differences in features between participant populations are statistically significant. In the subsequent chapters, we present the median difference and the 95% confidence interval (CI) for the differences, along with the p-value of the Kruskal-Wallis test.

[1] https://github.com/mne-tools/mne-python

Chapter 4

Audio Behavioral Analysis for Mental Disorders

In this chapter, a comprehensive analysis of audio-based features is conducted for diagnosis and assessment of schizophrenia and depression. First, the feature extraction from audio recordings is introduced. Then, this chapter presents the results on the classification of three different participant groups (classification tasks) and various of subjective assessment scores (prediction tasks) using verbal features (V), speech-based non-verbal features (N), and verbal and non-verbal features combined (VN), respectively. Lastly, the most salient features in paired classification tasks is presented to aid the understanding for psychiatrists.

4.1 Feature Extraction

In this thesis, both verbal and non-verbal feature sets were extracted from schizophrenia, depression, and healthy controls' recordings. The verbal cues are computed by the Linguistic Inquiry and Word Count (LIWC) method [113], Diction software [114], Latent Dirichlet Allocation (LDA) [123], and the Document to vector (Doc2Vec) method [124]. The non-verbal feature set includes prosodic features extracted by the Open-Source Media Interpretation by Large feature-space Extraction (openSMILE) toolbox [248], prosodic, articulate, and phonetic features computed by the DisVoice toolbox [249], in addition to conversational features [35]. The following briefly reviews these linguistic and non-verbal features. First,

the following sections explain the speaker classification in detail, which is the basic procedure for extracting participant speech.

4.1.1 Data Preprocessing

Before analyzing the recordings, two pre-processing steps were conducted. First, the segments recorded during the installation and removal of the recording equipment were manually removed. Second, the Audacity toolkit was applied to each channel to reduce the noise by typically 6dB. The noise statistics were automatically extracted from manually selected noisy segments, and the algorithm is described in the Audacity wiki [1].

4.1.2 Speaker Diarization

The speech of participants and the psychiatrists were recorded on separate channels, as mentioned earlier. Nevertheless, there is still some interference from the psychiatrist's channel onto the participant channel. In particular, automated speaker diarization was applied to remove the psychiatrists' voices from the participant channel. As illustrate in Figure 4.6, a Hidden Markov Model was applied to extract binary sequences from both audio channels to identify *who* is speaking and *when* (0: not speaking; 1: speaking). Channel 1 and 2 are the original signals, from which the binary sequences 1 and 2 were derived, indicating when the psychiatrists and participant respectively are speaking.

In order to obtain cohesive speech segments for speech recognition, one-dimensional erosion and dilation were applied to the binary sequence of the participant [250] (shown at the bottom of the Figure 4.6). Firstly, the binary sequence 2 was dilated by a one-second structuring element, which filled up small gaps (less than 1s) in one speech segment without filling up the adjacent two sentences. Next, the binary sequence 2 was eroded and dilated by a two-second structuring element. These steps reduce the noise and incorrect automated transcriptions. Finally, the filtered speech signal was obtained, which contains mostly speech from the participant, by multiplying the participant audio channel (channel 2) with the binary sequence associated with the participant (sequence 2).

[1] https://wiki.audacityteam.org/wiki/How_Audacity_Noise_Reduction_Works

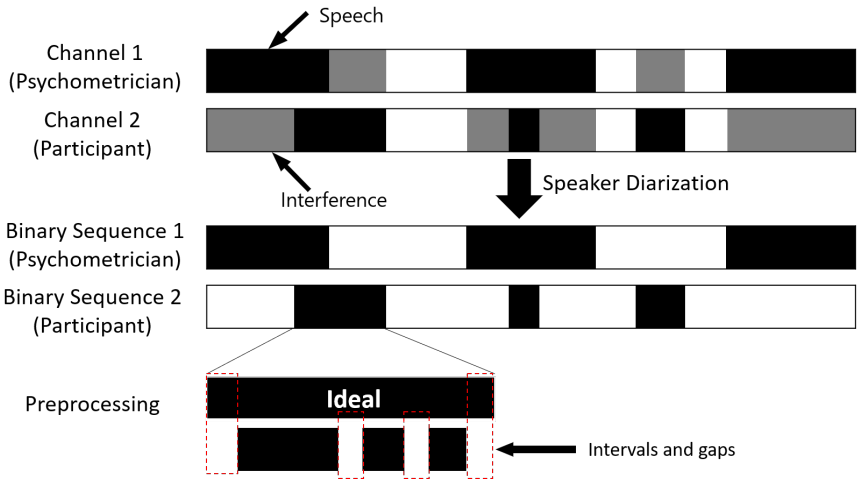


FIGURE 4.1: Illustration of speaker diarization.

4.1.3 Speech Recognition

After extracting the participant's speech, the Kaldi speech recognition toolkit was applied to automatically transcribe the participant's speech into text files. More specifically, the pre-trained ASpIRE Chain model [251] was utilized for automated transcriptions. This model is a DNN-HMM model, combining a deep neural network (DNN) with a hidden Markov model (HMM), pre-trained on Fisher English recordings, augmented with impulse responses and noises to create a multi-condition training dataset. To validate the speech recognition performance on our dataset, the first 5 minutes of 6 random audio samples were manually transcribed. The transcription accuracy of the ASpIRE Chain model on those 5-minute segments is about 52%. Although Kaldi's ASR model does not perform perfectly on these recordings, the verbal cues, such as the distribution of relative word frequency, extracted from automated transcriptions are similar to manual transcriptions, which is reported in Section 4.3 in detail.

4.1.4 Verbal Features

Linguistic features were extracted through the bag-of-words models LIWC 2015 [113], and Diction 7.0 software [114], where both of them represent texts by describing the occurrence of words within a document. The LIWC features comprise the word counts for 77 categories, including 21 linguistic dimensions counts: function words, common verbs, adjectives, etc.; 40 categories related to psychological

processes: words related to affect, sociality, cognition, perception, drive, biological processes, time orientations, and personal concerns, 6 informal Language markers: assents, fillers, swear words, question marks, netspeak, and all informal words; 7 personal concern categories: work, home, leisure activities, etc.; and 3 general text metrics: words counts in LIWC dictionary (dic), and the number of unique words (unique), words with more than six letters (sixltr). Similarly, Diction 7.0 generated 5 semantic features (Activity, Optimism, Certainty, Realism, and Commonality), 35 sub-features that form these semantic features, and 2 text metrics, i.e., number of unique words and average word size. Finally, the LIWC category counts and Diction sub-features were normalized by the total number of words.

Apart from word-based tools, transcriptions were also converted into vector space employing two unsupervised models: latent Dirichlet allocation (LDA) [123], and the Doc2Vec [124]. The LDA is a statistical model used to identify different topics of documents, where each document is modeled as a multinomial distribution of topics, and each topic is modeled as a multinomial distribution of words. It automatically generates the categories instead of manually determined. First, the top 100 topics were generated by LDA from transcripts of the participants' speech in the interviews. Next, the number of 50 most frequent words associated with those topics were counted and then normalized by the total number of words. The resulting normalized counts are treated as features for classification.

Moreover, the Doc2Vec model was leveraged to generate a document vector of transcriptions. Doc2Vec is a method that is used to generate representation vectors out of an article. Specifically, the document vectors were create by using the Distributed Memory of Paragraph Vector (PV-DM) algorithm [124] implemented in the Gensim library [252]. PV-DM algorithm let the model randomly sample consecutive words from a paragraph and predict a center word from the randomly sampled set of words, where the length of the document vector was set to 100. Both the LDA and Doc2Vec models were trained on the text files in the training set in each cross-validation (CV) loop with predefined epoch and random seed, while the document vectors were extracted from the text files in the test set by applying the trained models.

4.1.5 Non-verbal Features

Besides analyzing the interviews' linguistic contents, this thesis also analyzed low-level acoustic and prosodic patterns in the participants' speech by applying the OpenSmile [248], and Disvoice toolkits [253].The conversational features were also considered in our work. We reviewed low-level descriptors (LLDs) in the openSMILE and DisVoice toolkits, summarized in Table 6.1.

TABLE 4.1: LLDs in the openSMILE and DisVoice toolkits.

LLDs Name (Abbreviation)	Description
Intensity/Energy	Power carried by the audio waves
Loudness	Normalised intensity raised to the power of 0.3
Mel-frequency cepstral coefficients (MFCCs)	12 Mel-frequency Cepstral Coefficients from 25 ms audio frames with 10ms sliding window
Pitch (F0)	Fundamental frequency computed from the cepstrum
Probability of Voice (ProbVoice)	Voicing probability computed from the auto-correlation function (ACF)
F0 envelope (F0env)	Envelope of the smoothed fundamental frequency
Line Spectral Frequencies (LSF1-8)	Line spectral frequencies computed from 8 linear predictive coefficients
Zero Crossing Rate (ZCR)	Zero-crossing rate of time signal (frame-based)
First formant Frequency (FF1)	Frequency of the first formant
Second formant Frequency (FF2)	Frequency of the second formant
22 Bark band energies (BBEs)	Spectral energy over the 1-22 bark scales
Voice Durision (VoiceDur)	Duration of voice segments
Pause Durision (PauseDur)	Duration of non-voice segments
Pause Rate (PauseRate)	Percentage of non-voice segments in whole signal
Logaritmic Energy (LogE)	Logarithmic scale of acoustical power
Jitter	Average absolute difference between the frequency of consecutive periods
Shimmer	Average absolute difference between the amplitudes of consecutive periods
Amplitude perturbation quotient (APQ)	Average difference between the amplitude of five preceding and successive pitch periods
Pitch perturbation quotient (PPQ)	Variability of the pitch period evaluated in five consecutive cycles

The openSMILE toolkit [248] is a modular and adjustable collection of acoustic features useful for signal processing and machine learning applications. Specifically, the 'emobase' configuration of openSMILE was selected to extract the following LLDs: intensity, loudness, 12 Mel-frequency Cepstral Coefficients (MFCCs), pitch (F0), probability of voicing, F0 envelope (F0env), 8 line spectral frequencies (LspFreq), and Zero-Crossing Rate. Moreover, the following functions are applied to the LLDs and their delta coefficients (Delta): minimum and maximum values and their relative position form input (minPos and maxPos), range, mean, 2 linear regression coefficients (linregc1-2), linear and quadratic error, STD, skewness, kurtosis, values in 3 quartiles (quartile), and 3 inter-quartile ranges (quartile1-3). Before computing the LLDs, the pause and silence from the participant's speech were first removed, resulting in a continuous speech signal without silences. Then,

the LLDs from these continuous speech signals were extracted via a 100ms sliding window with no overlap. Finally, the mean, standard deviation (STD), minimum, and maximum of the LLDs values were calculated across all segments.

The DisVoice toolkit was also applied to the speech signals, which was first developed specifically for quantifying speech deficits of Parkinson patients [253]. The DisVoice toolkit provides articulation, prosody, and phonation features. The articulation features include the mean, STD, skewness, and kurtosis of the following speech measures: the first formant frequency (FF1), the second formant frequency (FF2), 22 bark band energies (BBEs), and 12 MFCCs with both *onset* (from unvoiced to voiced) and *offset* (from voiced to unvoiced) transitions, where this thesis also measured the first and second derivative of these features (e.g., DMFCC and DDMFCC). The prosody features include duration-based, F0-based, and energy-based measures. In the following, those three types of components were briefly described. The duration-based features comprise the mean, STD, minimum, and maximum duration of the voiced segments and pauses (VoiceDur and PauseDur) and the pause rate (number of pauses per second). The F0-based features consist of the mean and STD of F0 in voiced segments and semitones and average tilt and tilt regularity of F0, while the energy-based features comprise the mean, STD, and maximum values of logarithmic energy (LogE), voiced and unvoiced energy regularity, and regression coefficients and average tile of energy contour. Lastly, phonation features were computed over the voiced segments, including the mean, STD, skewness, and kurtosis of jitter, shimmer, amplitude perturbation quotient (APQ), pitch perturbation quotient (PPQ), logarithmic energy (LogVE), and the first and second derivative of F0 (DF0 and DDF0).

The interactions between participants and psychiatrists were also assessed in this study, similarly to in our early research [87]. A total of 14 conversational features were calculated from the speech of the participant and psychiatrist, extracted by speaker diarization: the number of short utterances (Interject), the average response time of participant (Response Time), average turn duration (Turn Duration), the percentage of speech (Speaking), the average duration of silence/pause (Speech Gap), the difference in the speaking percentages (Difference Speaking), the difference of natural turns (Difference Turn), word count per second (Speaking Rate), percentage of no speaking (Mutual Silence), percentage of duration when

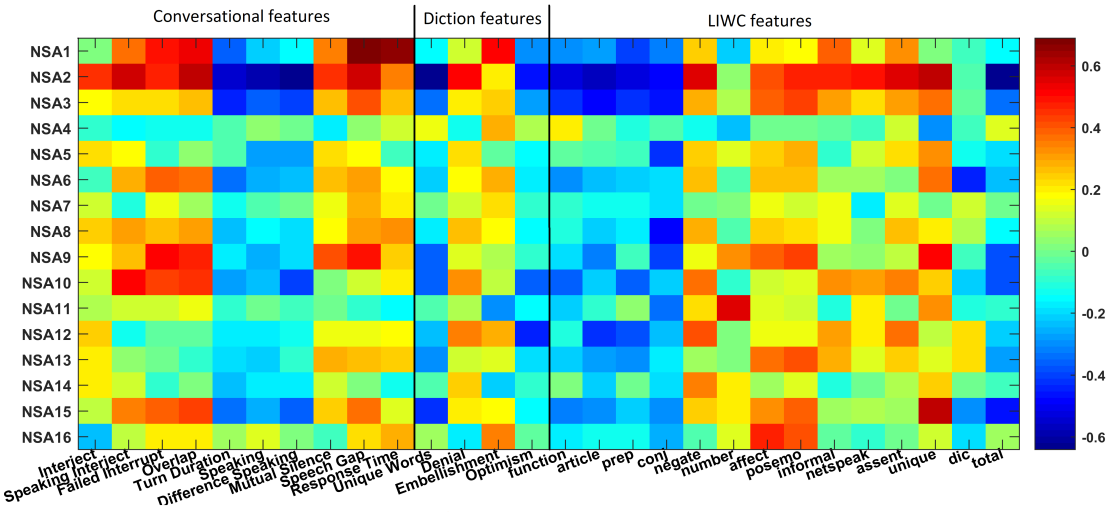
both speakers are speaking (Overlap), number of failed interrupts (Failed Interrupt), number of short utterances when another speaker is speaking (Speaking Interject), and the number of turns without interruption (Natural Turn). Some of those dynamic measurements are illustrated in Figure 4.2.

FIGURE 4.2: Illustration of the conversational cues. There is a bar for each of the two speakers, where a black (white) area indicates that the person is speaking (silent).

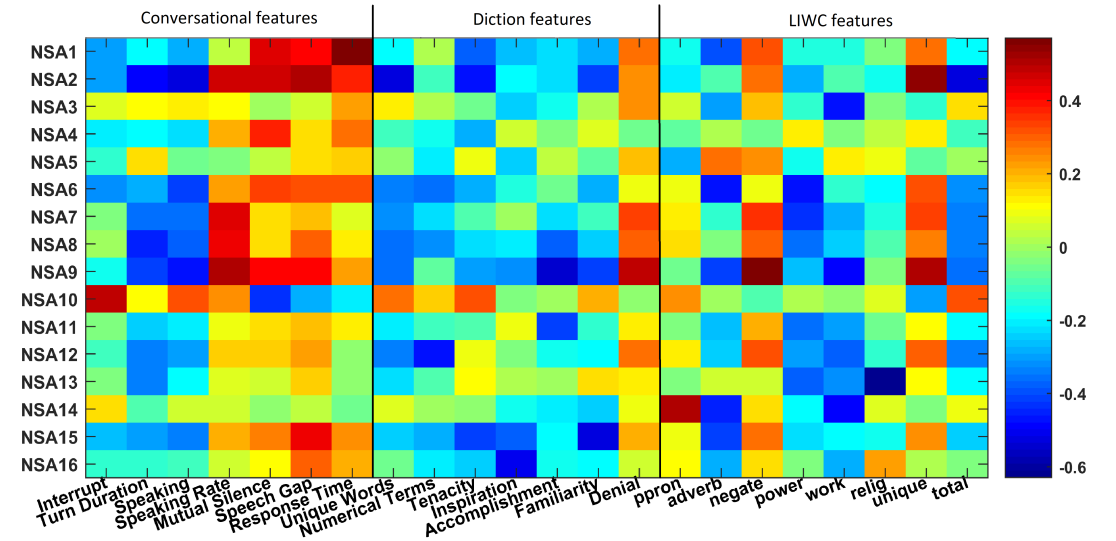
4.2 Correlation Analysis

To provide interpretable clinical useful information, the Pearson correlation was evaluated between our conversational and linguistic features (LIWC and Diction) and the subjective ratings (NSA-16) given by psychologists. It is noted that the correlation coefficients were only calculated for schizophrenic and depressed patients, as the NSA questionnaire was not designed for healthy controls. For conciseness, only features with a correlation coefficient of at least ± 0.45 are plotted in Figure 4.3. In schizophrenia and depression, there was a strong positive correlation between conversational features (mutual silence, speech gaps, and response time) and NSA1 (prolonged time to respond) and NSA2 (restricted speech volume), while features like turn duration and speaking were negatively correlated with NSA2. In addition, for linguistic features extracted from automated transcriptions, the total number of words and the number of unique words extracted by Diction (no normalization) were negatively correlated with the value of NSA2, however, the number of unique words extracted by LIWC (normalization by the total number of words) was positively correlated with NSA2. Besides, the proportion of function words in the total number of words, such as articles (e.g., a, an, the), prepositions (e.g., to,

with, above), and conjunctions (e.g., and, but, whereas), is negatively correlated with the value of NSA2 specific for schizophrenia.



(a) Schizophrenia.



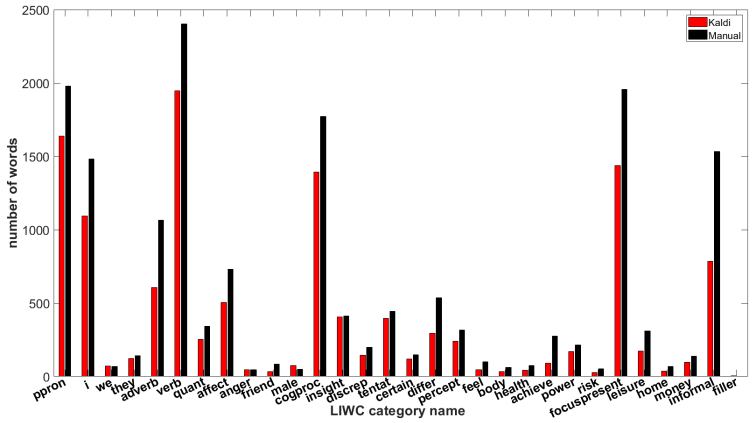
(b) Depression.

FIGURE 4.3: Correlation coefficients of NSA-16 scores with Conversational, LIWC, and Diction features for schizophrenia and depression respectively.

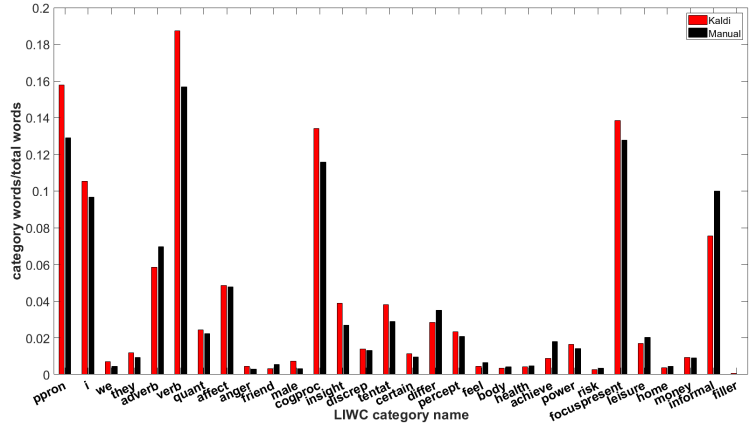
4.3 Manual vs Auto Transcriptions

To evaluate the Kaldi toolkit, an undergraduate student from NTU manually transcribed the entire recordings in the first session of Study A into texts. The total

word counts of each LIWC category between these manual transcriptions and corresponding Kaldi's transcriptions then were compared. The histograms of absolute and relative word frequencies are presented in Figure 4.4a and Figure 4.4b respectively, where only 20 random categories were presented. It shows that automatic transcriptions lose about one-fifth to one-third of words, but the frequency of the relative words of Kaldi's transcriptions is almost in line with the actual distribution.



(A) Absolute word counts of LIWC categories of manual and Kaldi transcriptions.



(B) Relative words frequency histogram of manual and Kaldi transcriptions.

FIGURE 4.4: The absolute and the relative words frequency histogram of word counts of LIWC categories of manual and Kaldi transcriptions.

Instead of comparing the total number of words in each category of manual and Kaldi transcriptions, the extracted linguistic features of each participant in both manual transcriptions and Kaldi transcriptions are compared in this thesis. The pairwise linear correlation coefficients between the linguistic features (LIWC and Diction features) extracted from the manual transcripts and those extracted from the Kaldi transcripts were plotted in Figure 4.5. The numbers in the X and Y axis

are the ID of the feature, where the first 77 features are LIWC features (ignore 'mark' feature in LIWC features as our texts do not contain question marks) and the left is 42 Diction features. it can be observed that most of the colors on the diagonal are dark red, which means most of the linguistic features extracted from Kaldi transcriptions are positively correlated with itself extracted from manual transcriptions. This further proves that reasonable results can be obtained by using the feature extracted from Kaldi transcriptions.

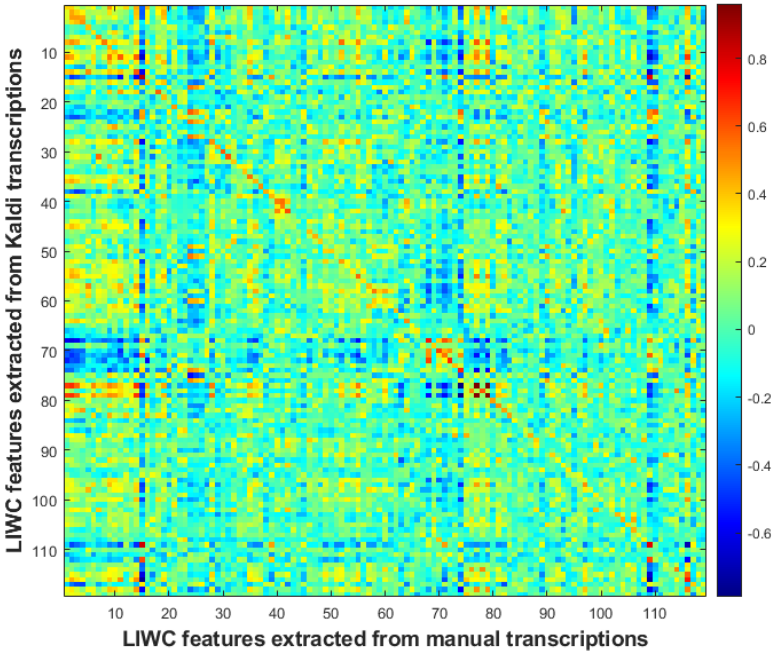


FIGURE 4.5: Correlation matrix plot of linguistic features extracted from Kaldi and Manual transcriptions.

4.4 Classification of Participants

As illustrated in Table 3.1, the audio recording from 48 patients with depression, 98 patients with schizophrenia, and 70 healthy controls were successfully recorded. As mentioned before, seven audio feature sets were extracted and late fusion ensemble learning with LOO-CV was applied to classify these three types of participants and each pair of them, and all the classification results are presented in Table 4.2. Notably, the objective audio features, verbal and non-verbal features combined, could differentiate patients with depression and healthy controls with a BAC of 80.6%,

TABLE 4.2: Results for automated classification of schizophrenia (S), depression (D), and healthy controls (H).

Task	Feature	CM			SEN	SPE	F1	MCC	AUPRC	ACC	BAC	MB
			Predicted									
			D	H								
D vs. H	V	D	35	13	0.729	0.729	0.731	0.451	0.799	0.729	0.729	0.593
		H	19	51								
	N	D	39	9	0.813	0.771	0.790	0.575	0.817	0.788	0.792	0.593
		H	16	54								
	VN	D	41	7	0.854	0.757	0.798	0.601	0.861	0.797	0.806	0.593
		H	17	53								
			H	**S**	**SEN**	**SPE**	**F1**	**MCC**	**AUPRC**	**ACC**	**BAC**	**MB**
S vs. H	V	H	49	21	0.867	0.700	0.795	0.580	0.847	0.798	0.784	0.583
		S	13	85								
	N	H	50	20	0.827	0.714	0.779	0.545	0.777	0.780	0.770	0.583
		S	17	81								
	VN	H	51	19	0.888	0.729	0.820	0.630	0.866	0.821	0.808	0.583
		S	11	87								
			D	**S**	**SEN**	**SPE**	**F1**	**MCC**	**AUPRC**	**ACC**	**BAC**	**MB**
D vs. S	V	D	36	12	0.694	0.750	0.720	0.419	0.788	0.712	0.722	0.671
		S	30	68								
	N	D	39	9	0.755	0.813	0.780	0.538	0.857	0.774	0.784	0.671
		S	24	74								
	VN	D	37	11	0.837	0.771	0.817	0.594	0.858	0.815	0.804	0.671
		S	16	82								
			H	**P**	**SEN**	**SPE**	**F1**	**MCC**	**AUPRC**	**ACC**	**BAC**	**MB**
P vs. H	V	H	47	23	0.671	0.822	0.774	0.488	0.826	0.773	0.747	0.676
		P	26	120								
	N	H	56	14	0.800	0.692	0.735	0.461	0.795	0.727	0.746	0.676
		P	45	101								
	VN	H	57	13	0.814	0.747	0.775	0.529	0.867	0.769	0.780	0.676
		P	37	109								
			D	**H**	**S**	**SEN**	**SPE**	**F1**	**MCC**	**AUPRC**	**ACC**	**BAC**	**MB**
D vs. S vs. H	V	D	31	11	6	0.541	0.864	0.602	0.404	0.691	0.593	0.605	0.454
		H	16	44	10								
		S	30	15	53								
	N	D	29	9	10	0.643	0.788	0.619	0.412	0.654	0.616	0.611	0.454
		H	14	41	15								
		S	17	18	63								
	VN	D	39	6	3	0.633	0.864	0.680	0.522	0.742	0.676	0.696	0.454
		H	12	45	13								
		S	24	12	62								

Note: We report here the classification results for verbal (V), non-verbal (N), and speech (VN) feature sets. Abbreviations: P=Patient; CM=Confusion Matrix; SEN=Sensitivity; SPE=Specificity; F1=F1-score; MCC=Matthews Correlation Coefficient; AUPRC=Area Under Precision-Recall Curve; ACC=Accuracy; BAC=Balanced Accuracy; MB=Majority Baseline.

significant improvement over the baseline accuracy of 59.3%. Non-verbal feature sets are considered to be the most important feature sets distinguishing between the

TABLE 4.3: Results for predicting the symptom severity using audio-based feature sets for all participants.

Symptoms	Scale	THR	Feature	CM		SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
				Predicted									
				L	H								
Negative Symptoms	NSA-Total	37.55	V	L 95	18	0.583	0.841	0.712	0.440	0.730	71.8%	71.2%	0.523
				H 43	60								
			N	L 86	27	0.709	0.761	0.736	0.471	0.784	73.6%	73.5%	0.523
				H 30	73								
			VN	L 87	26	0.738	0.770	0.755	0.508	0.784	75.5%	75.4%	0.523
				H 27	76								
	NSA-RS: Restricted speech	2.69	V	L 78	25	0.673	0.757	0.713	0.430	0.711	71.3%	71.5%	0.523
				H 37	76								
			N	L 72	31	0.770	0.699	0.736	0.470	0.766	73.6%	73.4%	0.523
				H 26	87								
			VN	L 84	19	0.673	0.816	0.740	0.491	0.752	74.1%	74.4%	0.523
				H 37	76								
	NSA-PQ: Poor quality of speech	2.52	V	L 92	21	0.592	0.814	0.704	0.418	0.733	70.8%	70.3%	0.523
				H 42	61								
			N	L 70	43	0.660	0.619	0.639	0.279	0.667	63.9%	64.0%	0.523
				H 35	68								
			VN	L 92	21	0.583	0.814	0.699	0.409	0.723	70.4%	69.8%	0.523
				H 43	60								
	NSA-AB: Affective blunting	5.06	V	L 73	34	0.761	0.682	0.722	0.445	0.744	72.2%	72.2%	0.505
				H 26	83								
			N	L 77	30	0.697	0.720	0.708	0.417	0.795	70.8%	70.8%	0.505
				H 33	76								
			VN	L 76	31	0.798	0.710	0.754	0.511	0.806	75.5%	75.4%	0.505
				H 22	87								
	NSA-AM: Amotivation	8.45	V	L 67	40	0.651	0.626	0.639	0.278	0.679	63.9%	63.9%	0.505
				H 38	71								
			N	L 71	36	0.633	0.664	0.648	0.297	0.652	64.8%	64.8%	0.505
				H 40	69								
			VN	L 79	28	0.587	0.738	0.660	0.329	0.704	66.2%	66.3%	0.505
				H 45	64								
Neuro-cognitive symptoms	BACS-VM	-0.30	V	L 70	38	0.574	0.648	0.611	0.223	0.638	61.1%	61.1%	0.500
				H 46	62								
			N	L 72	36	0.593	0.667	0.629	0.260	0.623	63.0%	63.0%	0.500
				H 44	64								
			VN	L 68	40	0.694	0.630	0.662	0.325	0.653	66.2%	66.2%	0.500
				H 33	75								
	BACS-DS	-0.23	V	L 55	49	0.705	0.529	0.617	0.238	0.619	62.0%	61.7%	0.519
				H 33	79								
			N	L 64	40	0.571	0.615	0.593	0.187	0.578	59.3%	59.3%	0.519
				H 48	64								
			VN	L 67	37	0.670	0.644	0.657	0.314	0.625	65.7%	65.7%	0.519
				H 37	75								
	BACS-TMT	-0.62	V	L 78	35	0.592	0.690	0.643	0.284	0.666	64.4%	64.1%	0.523
				H 42	61								
			N	L 74	39	0.709	0.655	0.681	0.363	0.672	68.1%	68.2%	0.523
				H 30	73								
			VN	L 85	28	0.583	0.752	0.669	0.340	0.707	67.1%	66.7%	0.523
				H 43	60								

Symptoms	Scale	THR	Feature	CM		SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
				Predicted									
				L	H								
Neuro-cognitive symptoms	BACS-SF	-0.27	V	L 66	44	0.632	0.600	0.616	0.232	0.669	61.6%	61.6%	0.509
				H 39	67								
			N	L 76	34	0.547	0.691	0.618	0.241	0.614	62.0%	61.9%	0.509
				H 48	58								
			VN	L 74	36	0.604	0.673	0.638	0.277	0.659	63.9%	63.8%	0.509
				H 42	64								
	BACS-SC	-0.81	V	L 70	38	0.722	0.648	0.685	0.371	0.722	68.5%	68.5%	0.500
				H 30	78								
			N	L 63	45	0.796	0.583	0.686	0.389	0.742	69.0%	69.0%	0.500
				H 22	86								
			VN	L 71	37	0.787	0.657	0.721	0.448	0.769	72.2%	72.2%	0.500
				H 23	85								
	BACS-ToL	0.26	V	L 73	42	0.545	0.635	0.592	0.180	0.594	59.3%	59.0%	0.532
				H 46	55								
			N	L 65	50	0.663	0.565	0.611	0.229	0.631	61.1%	61.4%	0.532
				H 34	67								
			VN	L 72	43	0.673	0.626	0.648	0.299	0.627	64.8%	65.0%	0.532
				H 33	68								
	BACS-Composite	-1.00	V	L 62	30	0.750	0.674	0.718	0.423	0.775	71.8%	71.2%	0.574
				H 31	93								
			N	L 62	30	0.718	0.674	0.700	0.389	0.767	69.9%	69.6%	0.574
				H 35	89								
			VN	L 71	21	0.726	0.772	0.747	0.492	0.821	74.5%	74.9%	0.574
				H 34	90								
	BACS-Composite	-2.00	V	L 28	16	0.756	0.636	0.751	0.337	0.826	73.1%	69.6%	0.796
				H 42	130								
			N	L 31	13	0.756	0.705	0.765	0.392	0.811	74.5%	73.0%	0.796
				H 42	130								
			VN	L 31	13	0.808	0.705	0.800	0.452	0.853	78.7%	75.6%	0.796
				H 33	139								
General psychiatric symptoms	BPRS-Total	24.00	V	L 60	15	0.674	0.800	0.724	0.451	0.774	71.8%	73.7%	0.653
				H 46	95								
			N	L 61	14	0.688	0.813	0.738	0.477	0.758	73.1%	75.1%	0.653
				H 44	97								
			VN	L 52	23	0.780	0.693	0.753	0.463	0.812	75.0%	73.7%	0.653
				H 31	110								
	BPRS-Total	32.00	V	L 88	50	0.769	0.638	0.691	0.391	0.728	68.5%	70.3%	0.639
				H 18	60								
			N	L 95	43	0.679	0.688	0.690	0.356	0.726	68.5%	68.4%	0.639
				H 25	53								
			VN	L 88	50	0.833	0.638	0.714	0.453	0.772	70.8%	73.6%	0.639
				H 13	65								
	BPRS-AFF: Affective	7.47	V	L 68	42	0.670	0.618	0.643	0.288	0.677	64.4%	64.4%	0.509
				H 35	71								
			N	L 53	57	0.698	0.482	0.583	0.184	0.595	58.8%	59.0%	0.509
				H 32	74								
			VN	L 67	43	0.745	0.609	0.675	0.357	0.674	67.6%	67.7%	0.509
				H 27	79								

Symptoms	Scale	THR	Feature	CM		SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
				Predicted									
				L	H								
General psychiatric symptoms	BPRS-POS: Positive	4.52	V	L 71	46	0.727	0.607	0.662	0.334	0.689	66.2%	66.7%	0.542
				H 27	72								
			N	L 80	37	0.677	0.684	0.681	0.360	0.706	68.1%	68.0%	0.542
				H 32	67								
			VN	L 77	40	0.737	0.658	0.695	0.395	0.735	69.4%	69.8%	0.542
				H 26	73								
	BPRS-NEG: Negative	6.04	V	L 75	28	0.779	0.728	0.754	0.508	0.747	75.5%	75.3%	0.523
				H 25	88								
			N	L 88	15	0.628	0.854	0.733	0.492	0.746	73.6%	74.1%	0.523
				H 42	71								
			VN	L 90	13	0.673	0.874	0.767	0.554	0.788	76.9%	77.3%	0.523
				H 37	76								
	BPRS-RES: Resistance	3.87	V	L 91	22	0.437	0.805	0.616	0.262	0.632	63.0%	62.1%	0.523
				H 58	45								
			N	L 82	31	0.544	0.726	0.636	0.274	0.670	63.9%	63.5%	0.523
				H 47	56								
			VN	L 73	40	0.592	0.646	0.620	0.238	0.676	62.0%	61.9%	0.523
				H 42	61								

Note: We report here the prediction results for verbal (V), non-verbal (N), and speech (VN) feature sets. Abbreviations: P=Patient; CM=Confusion Matrix; SEN=Sensitivity; SPE=Specificity; F1=F1-score; MCC=Matthews Correlation Coefficient; AUPRC=Area Under Precision-Recall Curve; ACC=Accuracy; BAC=Balanced Accuracy; MB=Majority Baseline.

depression and healthy groups (BAC=79.2%). Besides, patients with schizophrenia could be differentiated from healthy people with an accuracy of 80.8% by using audio-based features. Similarly, a multi-category classification task on depression vs. schizophrenia vs. healthy (BAC=69.6%) and pairwise classifications: depression vs. schizophrenia (BAC=80.4%) and patients vs. healthy (BAC=78.0%). These results indicate that every two kinds of participants exhibit significant differences in speech contents.

4.5 Prediction of Symptom Severity

As our analyses revealed strong discrimination between patients and healthy controls, the diagnosis targets (e.g., schizophrenia or depression) were not only classified, but also the assessment scores was predicted aiming for long-term monitoring. One obejetive of this thesis is to differentiate the negative symptoms assessed by NSA-16, neurocognitive symptoms assessed by BACS, general psychiatric symptoms assessed by BPRS, and social cognitive deficits measured by TASIT-P3. As

described in Section 3.5, each assessment score was categorized into 2 classes, class *High* or class *Low*, by a cut-off value. Next, various clinical scores are classified for schizophrenia patients, depression patients, and healthy controls combined. For each classification task, the late fusing results of verbal, nonverbal feature sets, as well as the combination of them are compared in the subsequent section. The prediction results for schizophrenia and depression were listed in Table 4.3. In the rest of this section, the prediction results of negative symptoms, neurocognitive symptoms, general psychiatric symptoms, social cognitive symptoms, and social cognitive symptoms were introduced respectively.

4.5.1 Negative Symptoms

According to the results, our methods performed well in predicting the negative symptom of patients with schizophrenia. It is interesting to note that, by soft voting the prediction outcomes of all verbal and non-verbal feature sets, a BAC of 75.4% was obtained for predicting NSA-Total from class *High* to class *Low*, outdoing the MB of about 20%. It can reasonably infer that our feature sets contain predictive information that is related to the overall negative symptom of schizophrenia patients.

Upon further investigation into the 4-Factor NSA scores, the analysis achieves the best BACs of 74.4% and 75.4% when classifying NSA-RS and NSA-AB of schizophrenia through late fusing all audio-based feature sets, where both of them got above 22% improvement than the majority baseline. We noted that using both language and non-verbal and facial feature sets are helpful for predicting NSA-RS, which may be because the quantity of speech reflected on the number of spoken words and duration of the speech. Moreover, it offers potential in predicting the NSA-AB using verbal and non-verbal characteristics, which may be due to abnormal word usage and monotonous speech. In addition, late fusing verbal feature sets seem to have predictive effects on the poor quality of speech (NSA-PQ, BAC=70.3%), whereas the prediction results of NSA-AM are less promising.

4.5.2 Neurocognitive Symptoms

As shown in Table 4.3, our methods performed well on predicting the negative symptom of patients with schizophrenia. The best classification result of BACS-Composite was obtained by soft voting the probability outputs of verbal and non-verbal feature sets, in which BAC reached 74.9% and 75.6% under the cut-off thresholds of -1 and -2 respectively. Since the BACS-Composite scale is split into three levels (normal, mild, and severe) as described in Section 3.5, these results indicate that our models were able to differentiate symptomatic and asymptomatic subjects as well as identify severe cases from the normal and mild cases. In particular, the verbal and non-verbal features can predict BACS-SC and BACS-TMT with a BAC of 72.2% and 66.7%, respectively.

4.5.3 General Psychiatric Symptoms

The general psychiatric symptoms assessed by BPRS contain a wide range of symptoms including negative, positive, cognitive, and depressive symptoms. The results of predictive measurement of general psychiatric symptoms (such as BPRS) tools can provide a comprehensive understanding of the severity of the patient's symptoms. Instead of first predicting the negative symptoms of schizophrenia and depression patients, the same scenario was applied for the general psychiatric symptoms assessed by BPRS. In this case, two cut-off thresholds (24 and 32) were used to segment BPRS-Total into normal, mild, and relatively severe. The BACs achieved on predicting BPRS-Total were 73.7% and 73.6% under these two thresholds, respectively. The best prediction results were achieved by late fusion of the prediction outputs of verbal and non-verbal feature sets. In addition, the BAC of prediction for the BPRS-NSA score reached 77.3%, which also confirms our previous results of predicting the severity of negative symptoms assessed by NSA-16.

4.6 Salient Features

Here, this thesis investigates what behavioral cues are significantly different in schizophrenia and depression patients than healthy controls. These cues enable the prediction of negative, cognitive, and general psychiatric symptoms. First, for

TABLE 4.4: Top 5 salient features for audio-based modality in paired classification tasks between schizophrenia (S), depression (D), and healthy control (H) groups.

Task	Linguistic Features	Non-verbal Features
$\overline{D}$ vs. H	**focuspast**^{a}, **feel**^{a}, adverb^{a}, **differ**^{a}, **i**^{a}, Certainty^{b}, Insistence^{b}, Collectives^{b}, **Exclusion**^{b}, Present Concern^{b}	Interject^{c}, **Response Time**^{c*}, Interrupt^{c*}, Speaking^{c}, Difference Turn^{c}, **lspFreq_sma[1]_max_min**^{d}, **lspFreq_sma_de[5]_max_min**^{d}, lspFreq_sma[5]_quartile2_max^{d}, **lspFreq_sma[1]_quartile2_min**^{d}, lspFreq_sma[5]_amean_max^{d}, mmlogE^{e*}, **stddurs**^{e*}, BBEoff11_std^{e*}, **PR**^{e*}, **BBEoff13_mean**^{e}
$\overline{S}$ vs. H	**percept**^{a**}, **death**^{a}, **focuspast**^{a}, **feel**^{a}, affiliation^{a}, **Communication**^{b}, **Aggression**^{b}, Collectives^{b}, **Concreteness**^{b}, Accomplishment^{b}	**Speech Gap**^{c}, Difference Turn^{c}, **Natural Turn**^{c}, Interrupt^{c}, Speaking^{c}, pcm_intensity_sma_de_minPos_mean^{d}, **lspFreq_sma_de[0]_max_min**^{d}, **mfcc_sma_de[11]_min_max**^{d}, F0env_sma_de_kurtosis_sd^{d}, **lspFreq_sma_de[3]_skewness_mean**^{d}, apq_sk^{e}, **BBEoff14_std**^{e}, **MFCCon9_mean**^{e}, **MFCCon8_std**^{e}, DMFCCoff12_mean^{e}
$\overline{D}$ vs. S	**sixltr**^{a}, **verb**^{a}, **auxverb**^{a}, **tentat**^{a}, **dic**^{a}, **Insistence**^{b}, **Certainty**^{b}, **Average Word Size**^{b}, **Tenacity**^{b}, **Hardship**^{b}	**Difference Turn**^{c}, **Speaking Rate**^{c}, Natural Turn^{c}, **Interject**^{c}, **Response Time**^{c**}, **lspFreq_sma[1]_max_min**^{d}, mfcc_sma[5]_minPos_mean^{d}, **mfcc_sma[2]_linregc1_mean**^{d}, F0_sma_de_quartile1_max^{d}, **lspFreq_sma_de[2]_min_max**^{d}, mmlogE^{e}, F0varsemi^{e}, **UVU**^{e}, **MFCCoff7_ku**^{e}, BBEoff13_mean^{e}

Feature set: a-LIWC; b-Diction; c-Conversational; d-OpenSmile; e-DisVoice; f-Affectiva; g-OpenFace; h-Opsis
Suffix: min-minimum; max-maximum; sk-skewness; ku-kurtosis; std-standard deviation; de-delta value; linregc-linear regression coef
**: p-value $\leq$ 0.005; *: p-value $\leq$ 0.05; Kruskal-Wallis tests with FDR correction.
bold feature: the average value of this feature is larger in the class with overline.

each behavioral cue, the feature importance was computed following the method described in Section 3.8. Next, the behavioral cues are ranked according to decreasing weights. The 5 behavioral cues with the largest weights and FDR post-correction p-values for verbal and non-verbal cues are presented in Table 4.4. Numerical results are presented with 95% confidence intervals (CI).

We noted that the FocusPast (0.0283, 95% CI - 0.0261 to 0.0306; P<0.05) and Certainty (43.09, 95% CI - 42.57 to 43.60; P<0.05) yield significant differences after FDR post-correction. The FocusPast quantifies the proportion of words related to the past (e.g., ago, did, past tense words, etc.). Certainty is a high-level linguistic feature that considers more semantic repetitions (the number of repetition words multiplies the sum of occurrences) and less self-focused words (e.g., I,

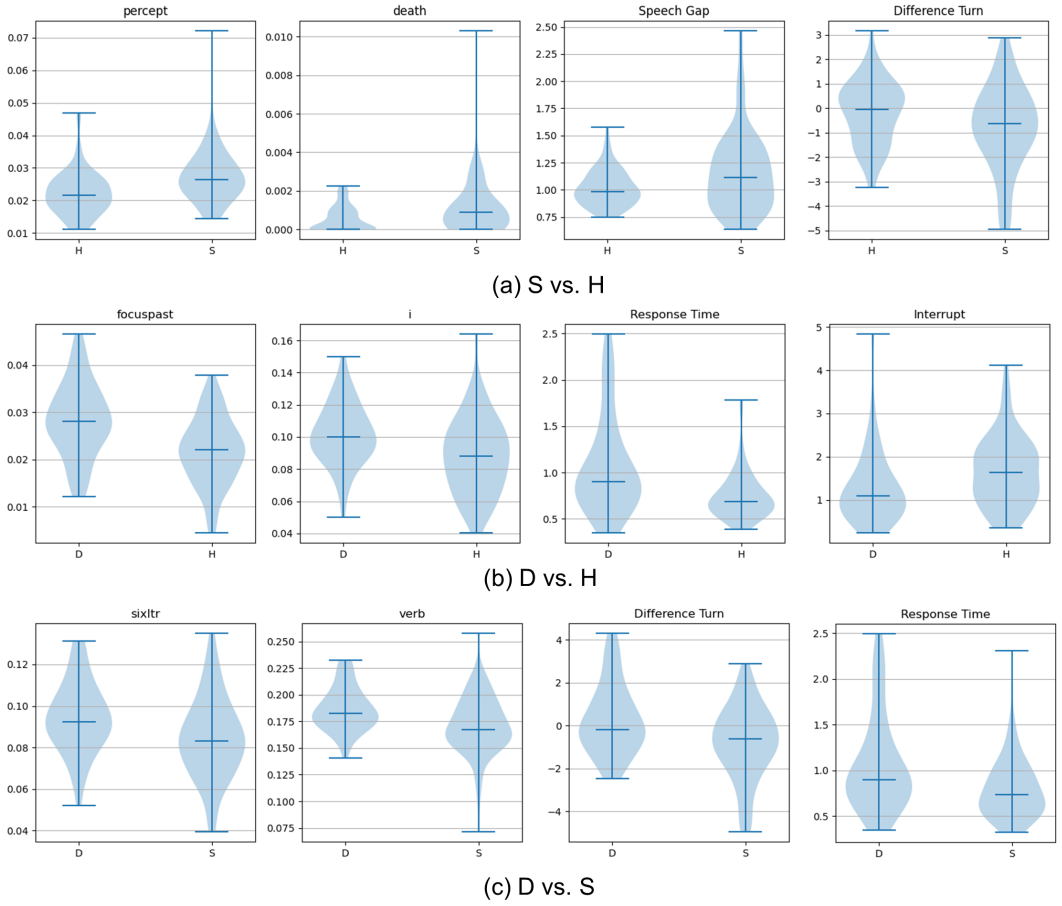


FIGURE 4.6: The violin plots of the top-ranked speech categories in paired classification tasks.

me, myself, etc.). The observation of more past focusing and self-reference from our study is similar to those underlined by Trifu and Smirnova [44, 133]. They also demonstrated that past tense words and self-focusing are frequent in the language of patients with depression. Our observations meet the self-focused theory of spoken language in depression [254]. Besides, the use of a larger proportion of past-focusing words of patients with depression may affect by the deficit in the integration of information from past experience [255]. As a result, patients with depression used more words to describe the past.

Through the non-verbal analysis, we noted that participants with depression had a significantly longer response time (1.07, 95% CI - 0.92 to 1.22; P<0.05). Our results are consistent with other investigations on time-related speech characteristics associated with depression [46, 160], which might be attributed to information

processing speed and psychomotor retardation [46]. Moreover, the feature ranking also revealed that LLDs related to LSF, F0, energy, MFCC, and their delta values had a significant role in the diagnosis of depression, where the F0, energy, and MFCC were also demonstrated by Jiang and Cummins [155, 256]. These differences may reflect the dull and monotonous speech of depression patients [257]. Besides, the number of interruptions and interjections are also significant on differentiation of patients with depression and healthy controls. It is also observed that a few depression patients rarely interrupt and interject the psychiatrist's conversation compared with healthy controls. However, to our best knowledge, there is no study that analyzed the interruption and interjection of patients with depression. Therefore, more studies of this nature are required.

Patients with schizophrenia also show significant language, speech, and facial expression abnormalities during the semi-structured interview. When looking at individual LIWC categories, we noted that the schizophrenic patients used a large proportion of words related to perception (0.0278, 95% CI - 0.0260 to 0.0295; $P<0.05$; e.g., see, feel, hear, etc.) and death (0.0012, 95% CI - 0.0009 to 0.0014; $P<0.05$; e.g., bury, coffin, kill, etc.) than healthy controls. The same pattern was also found in our previous study on a relatively small cohort [117]. Besides, Birnbaum et al. found that, in social media, the use of words in the death category of LIWC could predict the relapse of psychosis [258]. Minor et al. observed that the number of perceptual words was weakly correlated with the metacognition and had no correlation with the overall symptoms assessed by PANSS for schizophrenia [118]. In this study, perceptual word use was not correlated with the PANSS total score (Pearson's R=0.02), and there is a weak correlation between the perceptual words and greater neurocognitive symptoms (Pearson's R=0.17) and negative symptoms (Pearson's R=0.20) on patients with schizophrenia. The participants with schizophrenia in our study are mild patients and seem to prefer describing more about what they have seen, heard, and felt than healthy controls. These findings may demonstrate a suggestion raised by Rezaii et al. that the words related to perception may indicate early stages of change in auditory perception [217].

Furthermore, we noted that patients with schizophrenia had a longer duration of silence/pause than healthy controls, and the percentage of speaking time and the number of turns per minute for some patients are significantly low. Although these conversational cues are not significant after the post-correction of p-values, our

observations are in good agreement with the results summarized by Parola and further support the potential of conversational cues as a marker of schizophrenia symptomatology [32]. Moreover, the features related to the MFCC, spectral frequency (e.g., LSF), pitch (e.g., F0 and F0 envelope), and energy (e.g., BBE) were significant since they represented at least two-thirds of top features in the OpenSMILE and DisVoice feature sets. Similarly, other studies have also reported that these non-verbal features played an important role in distinguishing schizophrenia patients from healthy controls [34, 56, 259].

Interestingly, the proposed system also captured various behaviors between depression and schizophrenia, further supporting the detection of the transdiagnostic differences across diseases outlined by RDoC [60]. For instance, depression used more proportion of verbs (0.1866, 95% CI - 0.1806 to 0.1925; P<0.05) and the word with more than six letters, especially auxiliary verbs (0.0898, 95% CI - 0.0862 to 0.0934; P<0.05). However, our results are not in agreement with Lott et al., who found the use of the finite verbs did not show significant differences in the comparisons of patients with schizophrenia with people suffering from major depression [57].

TABLE 4.5: Summarized results for automated classification of schizophrenia, depression, and healthy controls using audio-based modalities.

Mode	Modality	Classification tasks				Prediction tasks					
						Negative symptoms		Cognitive symptoms		General psychiatric symptoms	
		S vs. H	D vs. H	D vs. S	DS vs. H	Global score	Total score	Normal vs. Mild	Mild vs. Severe	Normal vs. Mild	Mild vs. Severe
Single modality	LIWC	0.758	0.730	0.679	0.725	0.679	0.656	0.660	0.682	0.712	0.719
	Diction	0.673	0.625	0.696	0.684	0.624	0.640	0.666	0.662	0.644	0.622
	LDA	0.751	0.682	0.691	0.696	0.651	0.654	0.711	0.677	0.669	0.626
	Doc2Vec	0.671	0.732	0.697	0.663	0.617	0.663	0.648	0.633	0.643	0.589
	Conversational	0.650	0.704	**0.809**	0.638	0.705	**0.754**	0.658	0.667	0.714	0.615
	DisVoice	0.707	0.694	0.705	0.684	0.725	0.708	0.668	0.670	0.636	0.669
	OpenSmile	0.696	0.715	0.722	0.669	0.632	0.643	0.666	0.680	0.650	0.615
Fusion of modalities	Verbal	0.784	0.729	0.722	0.747	0.687	0.712	0.712	0.696	0.737	0.703
	Non-verbal	0.771	0.792	0.784	0.746	0.719	0.735	0.696	0.730	**0.751**	0.684
	Speech	**0.809**	**0.806**	0.804	**0.780**	**0.752**	**0.754**	**0.749**	**0.756**	0.737	**0.736**

Verbal modality: LIWC, Diction, LDA, and Doc2Vec.
Non-verbal modality: Conversational, DisVoice, and OpenSmile.
Speech modality: Verbal and Nonverbal modalities.

4.7 Conclusion

This study is cooperated with IMH Singapore and collected audio interview recordings from 103 schizophrenia patients, 50 depression patients, and 75 healthy controls. Speaker diarization, speech recognition, speech feature extraction, and ensemble learning with LOOCV were employed to automatically analyze both verbal and nonverbal analysis of people with schizophrenia and depression. This approach could further be applied in automated assessment and diagnosis of mentally ill patients. For better understanding the core results of this chapter, we summarized the classification results and prediction results in Table 4.5.

Many exciting results and interesting trends were observed in our analysis. The results indicate that, by using the objective verbal and non-verbal features extracted from automated transcriptions and audio itself, this thesis could provide acceptable predictions about the severity of the negative symptoms of schizophrenia patients. Besides, this thesis was also able to achieve high accuracy in the classification of different types of participants, which also has 11% to 22% improvement compared with the classification baseline (see Table 4.2). These results are promising and present an important step towards our overall goal of creating automated systems to aid clinical diagnosis and understanding of schizophrenia and depression.

Chapter 5

Audio-Visual Behavioral Analysis for Mental Disorders

Besides speech signals, this thesis also examines the video-based features, such as body movement and the affective expression on the patient's face. This chapter aims to analyze the audio and video cues as a whole and link these cues to their clinical status and conditions. This chapter first summarizes those video-based behavioral cues that were used in this study. This chapter then exhibits the classification and prediction results based on these video-based features and combined with the audio-based features outlined in Chapter 4. Specifically, the following 7 feature sets are considered: Verbal (V), Non-Verbal (N), Facial Expression (F), Body Movement (B), Verbal and Non-verbal (VN), Verbal, Non-verbal, and Facial Expression (VNF) feature sets, in addition to the combination of all individual feature sets (VNFB). Finally, the salient video-based features in the classification tasks and the cross-site validation results on the data of two studies are presented.

5.1 Feature Extraction

5.1.1 Facial Expression Features

Three different toolkits were applied to compute facial features: Affectiva, Opsis, and OpenFace. In each case, the entire video recordings of the interviews were processed. In other words, no specific episodes or events during the clinical interviews

were selected, but the full videos were analyzed instead. The facial expressions were illustrated in Figure 5.1. In the following, the facial expression cues considered in this study were summarized.

FIGURE 5.1: Facial expressions captured by Affectiva and OpenFace toolkits.

The Affectiva toolkit [260] calculates the probability value of 7 emotions (Anger, Contempt, Disgust, Fear, Joy, Sadness, and Surprise), 20 facial motions (e.g., MouthOpen, CheekRaise, NoseWrinkle, ChinRaise, EyeClosure, LipStretch, Smirk, etc.), and 13 emojis with (e.g., Laughing, Smiley, Wink, Relaxed, Scream, Stuck-OutTongue, etc.). In addition, the Opsis toolkit[1] quantifies emotions in a 3-dimensional continuous space: Arousal (passive vs. energetic), Valence (negative vs. positive), and Intensity (difference from neutral). Besides these three emotional metrics, 3 head postures (Roll, Pitch, and Yaw angles) and 1 eye openness feature (Lambda) are also measured by the Opsis toolkit. Finally, the OpenFace toolkit [261] was applied to quantify facial expressions as well. This toolkit automatically captures

[1]http://www.opsis.sg/

2 eye-gaze directions in world coordinates (GazeAngle_x for vertical axis; GazeAngle_y for horizontal axis), 6 rigid shape parameters (scale, rotation, and translation terms, denoted by P_scale, P_rx and P_ry, and P_rz, P_tx, and P_ty, respectively), 34 non-rigid shape parameters (NSP0 to NSP33), the regression intensity of 17 Facial Action Units (AU01_reg to AU17_reg), and the classification values of these AUs in a binary format (AU01_clf to AU17_clf).

The differences of the features across consecutive frames were also calculated (referred to as delta values), indicating how much the features change over time. Next, the statistical measures of those features were then computed across the entire length of the videos. For instance, the mean, minimum, maximum, median, skewness, and kurtosis of all Affectiva and Opsis features (except the three head postures) and their delta values were calculated. In addition, the percentage of Affectiva scores above a threshold of 10 (maximum is 100) was also included into the Affectiva feature set to measure the duration of emotions and facial expressions. Finally, for OpenFace features extracted across consecutive frames, this study calculated the mean of AUs classification values and the mean, minimum, maximum, median, skew, and kurtosis values of other OpenFace features (face shape parameters and gaze direction).

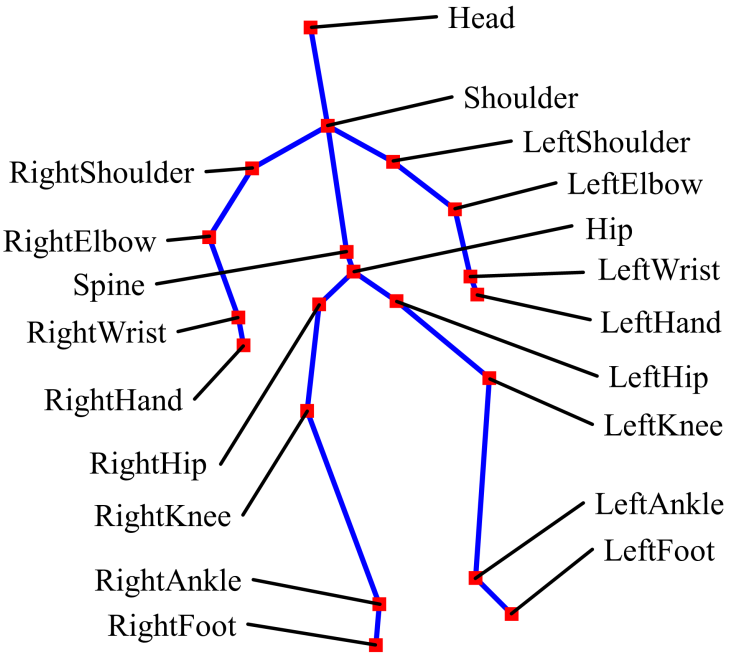


FIGURE 5.2: Body joints captured by Microsoft Kinect.

5.1.2 Body Movement Features

The Microsoft Kinects was implemented to automatically extract keletal points from depth recordings in the experiment. The Kinect v1 and Kinect v2 capture the 3D positions of 20 and 25 joints, and they were used in the first study and the second study respectively. To maintain consistency across the first study (Kinect v1) and the second study (Kinect v2), only the 20 joints captured by both Kinect v1 and Kinect v2 were analyzed. The names of those joints are shown in Figure 5.2. Moreover, a median filter with a one-second sliding window was first applied to remove spurious noise. The linear speed (LiteSpeed) of all 20 joints was then measured by calculating the differences between adjacent frames, and their mean and STD were computed. Apart from the linear velocity of the joints, this work also evaluated the angular speed (AngSpeed) and acceleration (AngAcc) of 6 body angles (left and right shoulder, elbow, and wrist joints). Similarly, the mean and STD for all angular speeds and accelerations were also calculated. Finally, a total of 64 features are fused as one body movement feature set.

5.2 Classification of Participants

All the behavioral cues extracted from audio and video recordings were used to classify 50 depression patients, 103 schizophrenic patients, and 75 healthy controls with LOO-CV (See Table 5.1). In addition, a multi-category classification task was performed on depression vs. schizophrenia vs. healthy (BAC=68.7%) and pairwise classifications: depression vs. healthy (BAC=82.3%), schizophrenia vs. healthy (BAC=82.3%), depression vs. schizophrenia (BAC=84.7%), and patients vs. healthy (BAC=79.8%). The highest classification results were obtained for all the above cases by fusing the prediction outputs from all feature sets (verbal, nonverbal, facial, and body movement).

5.2.1 Classifier Selection

To compare different state-of-art classifiers on each of the pipelines towards depression and schizophrenia recognition, I made an in-depth comparison and summarized the balanced accuracy for each classifier and feature set, as shown in Table 5.2. It is

TABLE 5.1: Results for automated classification of schizophrenia (S), depression (D), and healthy controls (H).

Task	**Feature**	**CM**			**SEN**	**SPE**	**F1**	**MCC**	**AUPRC**	**ACC**	**BAC**	**MB**
			Predicted									
			D	**H**								
D vs. H	F	D	23	19	0.548	0.778	0.662	0.335	0.714	0.667	0.663	0.517
		H	10	35								
	B	D	29	13	0.690	0.682	0.689	0.364	0.696	0.685	0.686	0.611
		H	21	45								
	VN	D	41	7	0.854	0.757	0.798	0.601	0.861	0.797	0.806	0.593
		H	17	53								
	VNF	D	41	9	0.820	0.784	0.800	0.594	0.865	0.798	0.802	0.597
		H	16	58								
	VNFB	D	37	13	0.740	0.907	0.838	0.663	0.879	0.840	0.823	0.600
		H	7	68								
			H	**S**	**SEN**	**SPE**	**F1**	**MCC**	**AUPRC**	**ACC**	**BAC**	**MB**
S vs. H	F	H	31	14	0.795	0.689	0.741	0.487	0.795	0.742	0.742	0.506
		S	9	35								
	B	H	40	26	0.696	0.606	0.659	0.301	0.698	0.658	0.651	0.582
		S	28	64								
	VN	H	51	19	0.888	0.729	0.820	0.630	0.866	0.821	0.808	0.583
		S	11	87								
	VNF	H	52	22	0.939	0.703	0.834	0.673	0.873	0.838	0.821	0.572
		S	6	93								
	VNFB	H	55	20	0.913	0.733	0.835	0.665	0.889	0.837	0.823	0.579
		S	9	94								
			D	**S**	**SEN**	**SPE**	**F1**	**MCC**	**AUPRC**	**ACC**	**BAC**	**MB**
D vs. S	F	D	30	12	0.773	0.714	0.744	0.488	0.774	0.744	0.744	0.512
		S	10	34								
	B	D	30	12	0.707	0.714	0.718	0.395	0.793	0.709	0.710	0.687
		S	27	65								
	VN	D	37	11	0.837	0.771	0.817	0.594	0.858	0.815	0.804	0.671
		S	16	82								
	VNF	D	43	7	0.808	0.860	0.829	0.640	0.893	0.826	0.834	0.664
		S	19	80								
	VNFB	D	41	9	0.874	0.820	0.858	0.681	0.905	0.856	0.847	0.673
		S	13	90								
			H	**P**	**SEN**	**SPE**	**F1**	**MCC**	**AUPRC**	**ACC**	**BAC**	**MB**
P vs. H	F	H	34	11	0.756	0.709	0.732	0.444	0.772	0.725	0.732	0.656
		P	25	61								
	B	H	31	35	0.470	0.746	0.654	0.217	0.684	0.655	0.608	0.670
		P	34	100								
	VN	H	57	13	0.814	0.747	0.775	0.529	0.867	0.769	0.780	0.676
		P	37	109								
	VNF	H	56	18	0.757	0.819	0.801	0.561	0.854	0.798	0.788	0.668
		P	27	122								
	VNFB	H	58	17	0.773	0.824	0.810	0.580	0.862	0.807	0.798	0.671
		P	27	126								

Task	Feature	CM				SEN	SPE	F1	MCC	AUPRC	ACC	BAC	MB
			Predicted										
			D	H	S								
D vs. S vs. H	F	D	21	8	13	0.568	0.690	0.543	0.314	0.615	0.542	0.541	0.344
		H	6	25	14								
		S	11	8	25								
	B	D	18	13	11	0.511	0.694	0.474	0.184	0.556	0.470	0.460	0.460
		H	15	29	22								
		S	18	27	47								
	VN	D	39	6	3	0.633	0.864	0.680	0.522	0.742	0.676	0.696	0.454
		H	12	45	13								
		S	24	12	62								
	VNF	D	34	11	5	0.657	0.839	0.663	0.485	0.762	0.659	0.662	0.444
		H	11	48	15								
		S	21	13	65								
	VNFB	D	35	10	5	0.680	0.840	0.687	0.519	0.780	0.684	0.687	0.452
		H	9	51	15								
		S	19	14	70								

Note: We report here the classification results for verbal (V), non-verbal (N), and speech (VN) feature sets. Abbreviations: P=Patient; CM=Confusion Matrix; SEN=Sensitivity; SPE=Specificity; F1=F1-score; MCC=Matthews Correlation Coefficient; AUPRC=Area Under Precision-Recall Curve; ACC=Accuracy; BAC=Balanced Accuracy; MB=Majority Baseline.

clear that no single classifier performed best in all classification tasks. Therefore, in this study, we late fused the prediction results of all five classifiers to obtain robust results.

TABLE 5.2: Balanced accuracy for automated classification of schizophrenia (S), depression (D), and healthy controls (H) using different classifiers and feature sets.

Task	Classifier	Verbal	NonVerbal	Facial	Speech	Speech+Facial	All
D vs. H	Logistic Regression	0.709	0.737	0.670	0.772	0.799	0.800
	SVM	0.727	0.764	0.640	**0.810**	0.785	0.797
	Gradient Boosting	0.733	0.732	0.657	0.795	0.778	0.800
	AdaBoost	0.656	0.675	0.679	0.744	0.772	0.723
	RandomForest	**0.763**	0.764	**0.710**	0.807	0.792	0.793
	Ensemble classifier	0.729	**0.792**	0.663	0.806	**0.802**	**0.823**
S vs. H	Logistic Regression	**0.792**	0.752	0.719	**0.817**	**0.846**	**0.861**
	SVM	0.771	0.702	0.695	0.789	0.786	0.790
	Gradient Boosting	0.755	0.673	0.732	0.784	0.777	0.789
	AdaBoost	0.714	0.670	**0.775**	0.735	0.765	0.777
	RandomForest	0.760	0.707	0.698	0.773	0.781	0.784
	Ensemble classifier	0.784	**0.770**	0.742	0.808	0.821	0.823
D vs. S	Logistic Regression	0.666	0.758	0.707	0.753	0.784	0.798
	SVM	0.706	0.774	0.756	0.780	0.819	0.833
	Gradient Boosting	**0.747**	**0.788**	0.661	0.799	0.789	0.813
	AdaBoost	0.705	0.712	**0.801**	0.752	0.829	0.832
	RandomForest	0.757	0.738	0.744	0.799	0.819	0.838
	Ensemble classifier	0.722	0.784	0.744	**0.804**	**0.834**	**0.847**

5.2.2 Fine-tune Classifier vs. Ensemble Classifier

To understand whether a well-tuned base classifier outperforms the ensemble of inferior base classifiers, we compared the classification results using our proposed ensemble classifier and the fine-tune SVM classifier, as shown in Table 5.3.

Technically, we compared the classification results using the proposed ensemble classifier and the SVM classifier with RBF kernel (SVM-RBF) and polynomial kernel (SVM-Poly). The cross-validation (CV) grid search was used to select the best hyperparameters of the classifier. The CV grid search is a method to search the hyper-parameter space for the best CV accuracy. we applied leave-one-out CV to validate the classification accuracy for the following results using SVM-RBF and SVM-Poly classifiers. In each CV loop, we employed a 10-fold CV grid search to find the best hyper-parameters to optimize the CV accuracy. I applied parameter optimization on the regularization parameter (C $\in$ [0.001,0.05,0.01,0.5,1,5,10]) and the kernel coefficient (gamma $\in$ [0.001,0.05,0.01,0.5,1,5,10]) for SVM-RBF and SVM-Poly classifiers, and degree (degree $\in$ [1,2,3,4,5])) for SVM-Poly classifier.

We found that the proposed ensemble classifier achieved the best performance in almost all classification tasks (such as S vs. H, D vs. S, and DS vs. H). For D vs. H, using a fine-tune SVM classifier with RBF core achieves better performance than using an integrated classifier. The ensemble classifier seems to be more versatile and robust in different classification tasks, and a well-tuned base classifier may also provide good classification results in some cases. One disadvantage of fine-tuning a classifier is that it is usually not time efficient. In the future, we will continue to explore whether the well-tuned base classifier will have good results on different classification and prediction tasks.

5.3 Prediction of Symptom Severity

5.3.1 Negative Symptoms

In addition to identifying the type of mental disorder, this thesis also attempted to use behavioral cues to infer the negative symptom severity in patients with schizophrenia and depression. The prediction results for predicting the four weighted

TABLE 5.3: Balanced accuracy for classification of schizophrenia (S), depression (D), and healthy controls (H) using ensemble classifier and fine-tune SVM classifier.

Task	Classifier	Verbal	Non-verbal	Facial	Speech	Speech + Facial	All
S vs. H	Ensemble classifier	**0.784**	**0.770**	**0.742**	**0.808**	0.821	**0.823**
	SVM-rbf	0.778	0.699	0.730	0.801	**0.825**	0.811
	SVM-poly	0.760	0.642	0.638	0.773	0.771	0.757
D vs. H	Ensemble classifier	0.729	0.792	0.663	0.806	0.802	0.823
	SVM-rbf	**0.788**	**0.915**	**0.906**	**0.919**	**0.926**	**0.927**
	SVM-poly	0.687	0.721	0.598	0.740	0.735	0.720
D vs. S	Ensemble classifier	0.722	0.784	**0.744**	**0.804**	**0.834**	**0.847**
	SVM-rbf	**0.763**	0.789	0.439	0.784	0.749	0.780
	SVM-poly	0.743	**0.800**	0.655	0.799	0.809	0.818
DS vs. H	Ensemble classifier	0.754	**0.761**	0.723	**0.787**	**0.802**	0.812
	SVM-rbf	**0.756**	0.647	**0.744**	0.741	0.754	**0.842**
	SVM-poly	0.739	0.691	0.729	0.765	0.765	0.766

factor scales derived from the NSA scores and the NSA total score are summarized in Table 5.4. The proposed methods consistently achieved good accuracy (BAC of 76.6%) in predicting the NSA-Total score (High vs. Low) for patients with schizophrenia. Similarly, the BAC was 76.0% when all three types of subjects were included. In addition, when predicting the factor scores of negative symptoms, the methods achieved better results for NSA-RS and NSA-AB than for NSA-PQ and NSA-AM. As shown in Table 5.4, under the sample "S" category, we note that the NSA-RS and NSA-AB achieve a BAC of 79.8% and 71.9% compared with NSA-PQ and NSA-AM of 63.9% and 69.9%. This difference implies that the objective behavioral cues are more correlated to the expression-related scales than motivation-related scales.

Instead of predicting the total score and factor scores of NSA-16, the prediction results of individual ratings of NSA-16 are detailed in Table A.1. Besides, the NSA1, NSA2, NSA6, and NSA15 indices could be generally well predicted through behavioral features across different participants. The BACs are observed to be significantly improved comparing with the baseline accuracy. NSA2 refers to the symptom of restricted speech quantity, which is highly correlated with the number of function words and the total number of words patients used during the interview. Additionally, our analysis also revealed that NSA 6 (Reduced modulation of intensity) and NSA15 (Reduced expressive gestures) related to the emotional and gestural expressions could also be reliably predicted. It is often observed that individual who speaks less also gesticulate less since hand gestures and body

TABLE 5.4: Results for automated prediction of the severity of the negative symptoms assessed by NSA16.

Samples	Scale	THR	Feature	CM			SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
					Predicted									
					L	H								
S	NSA-RS	3.18	N	L	41	12	0.822	0.774	0.796	0.594	0.818	0.796	0.798	0.541
				H	8	37								
	NSA-PQ	3.14	VN	L	20	21	0.789	0.488	0.655	0.292	0.597	0.663	0.639	0.582
				H	12	45								
	NSA-AB	6.37	VNFB	L	37	14	0.712	0.725	0.718	0.437	0.713	0.718	0.719	0.505
				H	15	37								
	NSA-AM	9.16	F	L	19	4	0.571	0.826	0.699	0.413	0.718	0.705	0.699	0.523
				H	9	12								
	NSA-Total	41.00	VN	L	39	9	0.740	0.813	0.775	0.553	0.752	0.776	0.776	0.510
				H	13	37								
D	NSA-RS	2.82	N	L	14	10	0.792	0.583	0.684	0.383	0.663	0.688	0.688	0.500
				H	5	19								
	NSA-PQ	3.36	V	L	10	12	0.846	0.455	0.652	0.330	0.629	0.667	0.650	0.542
				H	4	22								
	NSA-AB	6.11	V	L	18	7	0.609	0.720	0.666	0.331	0.593	0.667	0.664	0.521
				H	9	14								
	NSA-AM	9.87	VN	L	17	5	0.577	0.773	0.665	0.353	0.599	0.667	0.675	0.542
				H	11	15								
	NSA-Total	41.00	N	L	16	9	0.652	0.640	0.646	0.292	0.557	0.646	0.646	0.521
				H	8	15								
DSH	NSA-RS	2.82	VNF	L	83	33	0.841	0.716	0.775	0.559	0.810	0.776	0.778	0.520
				H	17	90								
	NSA-PQ	2.55	V	L	92	21	0.592	0.814	0.704	0.418	0.733	0.708	0.703	0.523
				H	42	61								
	NSA-AB	5.30	VNF	L	91	21	0.748	0.813	0.780	0.562	0.837	0.780	0.780	0.502
				H	28	83								
	NSA-AM	8.45	VNF	L	84	28	0.604	0.750	0.675	0.358	0.709	0.677	0.677	0.502
				H	44	67								
	NSA-Total	38.00	VNF	L	84	33	0.802	0.718	0.758	0.520	0.806	0.758	0.760	0.525
				H	21	85								

Note: The scores are divided into Class High (H) and Class Low (L) by a cut-off threshold (THR). The THRs of the 4-factor scores of NSA and NSA-Total are set as their median. Best prediction results for verbal (V), non-verbal (N), facial expression(F), and body movement (B) feature sets are presented. NSA=16-item Negative Symptoms Assessment; RS=Restricted Speech; PQ=Poor Quality of Speech; AB=Affective Blunting; AM=Amotivation; CM=Confusion Matrix; SEN=Sensitivity; SPE=Specificity; AUPRC=Area Under Precision-Recall Curve; ACC=Accuracy; BAC=Balanced Accuracy; MB=Majority Baseline.

movements often accompany speech. These NSA symptoms are highly interrelated and underlie the impairments of cognitive and emotional processes in individuals suffering from negative symptoms of schizophrenia.

TABLE 5.5: Results for automated prediction of the severity of the neurocognitive symptoms assessed by BACS.

Samples	Scale	THR	Feature	CM		SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
				Predicted									
				L	H								
S	BACS-VM	-0.79	VNFB	L 30	23	0.600	0.566	0.583	0.166	0.568	0.583	0.583	0.515
				H 20	30								
	BACS-DS	-0.95	B	L 26	21	0.600	0.553	0.576	0.153	0.568	0.576	0.577	0.511
				H 18	27								
	BACS-TMT	-1.26	VN	L 41	6	0.647	0.872	0.753	0.530	0.739	0.755	0.760	0.520
				H 18	33								
	BACS-SF	-1.01	N	L 28	21	0.653	0.571	0.612	0.225	0.595	0.612	0.612	0.500
				H 17	32								
	BACS-SC	-1.52	N	L 30	18	0.580	0.625	0.602	0.205	0.533	0.602	0.603	0.510
				H 21	29								
	BACS-ToL	-0.06	VN	L 39	12	0.404	0.765	0.578	0.182	0.510	0.592	0.584	0.520
				H 28	19								
	BACS-Composite	-1.00	VNF	L 52	14	0.667	0.788	0.750	0.445	0.726	0.747	0.727	0.667
				H 11	22								
	BACS-Composite	-2.00	F	L 12	3	0.621	0.800	0.690	0.399	0.733	0.682	0.710	0.659
				H 11	18								
D	BACS-VM	-0.01	N	L 17	5	0.692	0.773	0.730	0.464	0.642	0.729	0.733	0.542
				H 8	18								
	BACS-DS	0.06	V	L 13	11	0.792	0.542	0.661	0.344	0.669	0.667	0.667	0.500
				H 5	19								
	BACS-TMT	-0.46	F	L 18	5	0.737	0.783	0.762	0.519	0.843	0.762	0.760	0.548
				H 5	14								
	BACS-SF	0.47	N	L 19	5	0.417	0.792	0.590	0.225	0.571	0.604	0.604	0.500
				H 14	10								
	BACS-SC	-0.20	F	L 18	2	0.545	0.900	0.706	0.472	0.631	0.714	0.723	0.524
				H 10	12								
	BACS-ToL	0.58	B	L 19	3	0.600	0.864	0.733	0.483	0.693	0.738	0.732	0.524
				H 8	12								
	BACS-Composite	-1.00	VNFB	L 12	1	0.703	0.923	0.775	0.551	0.809	0.760	0.813	0.740
				H 11	26								

TABLE 5.5: Results for automated prediction of the severity of the neurocognitive symptoms assessed by BACS.

Samples	Scale	THR	Feature	CM		SEN	SPE	F1	MCC	AUPRC	ACC	BAC	MB
				Predicted									
				L	H								
DSH	BACS-VM	-0.30	VNF	L 72	38	0.699	0.655	0.677	0.354	0.655	0.677	0.677	0.507
				H 34	79								
	BACS-DS	-0.23	VN	L 67	37	0.670	0.644	0.657	0.314	0.625	0.657	0.657	0.519
				H 37	75								
	BACS-TMT	-0.62	N	L 74	39	0.709	0.655	0.681	0.363	0.672	0.681	0.682	0.523
				H 30	73								
	BACS-SF	-0.27	VN	L 74	36	0.604	0.673	0.638	0.277	0.659	0.639	0.638	0.509
				H 42	64								
	BACS-SC	-0.81	VNFB	L 89	23	0.690	0.795	0.741	0.486	0.797	0.741	0.742	0.509
				H 36	80								
	BACS-ToL	0.14	VNFB	L 77	47	0.712	0.621	0.663	0.332	0.644	0.662	0.666	0.544
				H 30	74								
	BACS-Composite	-1.00	VNFB	L 73	23	0.818	0.760	0.794	0.578	0.822	0.794	0.789	0.579
				H 24	108								
	BACS-Composite	-2.00	VN	L 31	13	0.808	0.705	0.800	0.452	0.853	0.787	0.756	0.796
				H 33	139								

Note: All BACS scores are the Z-Scores, which are standardized on an external Singapore dataset.69 The scores are divided into Class High (H) and Class Low (L) by a cut-off threshold (THR). The THRs of the six domain scores of BACS are set as their median. The BACS-Composite THR for normal and mild illness is set to -1. Besides, the BACS-Composite THR for mild and severe illness is set to -2. Best prediction results for verbal (V), non-verbal (N), facial expression(F), and body movement (B) feature sets are presented. RS=Restricted Speech; PQ=Poor Quality of Speech; AB=Affective Blunting; AM=Amotivation; CM=Confusion Matrix; SEN=Sensitivity; SPE=Specificity; AUPRC=Area Under Precision-Recall Curve; ACC=Accuracy; BAC=Balanced Accuracy; MB=Majority Baseline.

5.3.2 Cognitive Symptoms

The behavioral audio-visual cues were also utilized to predict cognitive symptoms measured by BACS (Table 5.5). The cut-off scores of -1 and -2 were chose to divide the composite score of BACS (BACS-Composite) into three levels: normal (schizophrenia=33, depression=37, healthy=62), mild (schizophrenia=33, depression=9, healthy=7), and severe (schizophrenia=37, depression=4, healthy=6). For detecting mild to severe cognitive symptoms (BACS-Composite<-1), as shown in Table 5.5, the balanced accuracies for patients with schizophrenia, depression, and all three types of subjects combined were 72.7%, 81.3%, and 78.9%, respectively. Similarly, for detecting severe cognitive symptoms (BACS-Composite<-2) in patients with schizophrenia and all three groups of subjects, the balanced accuracies were 71.0% and 75.6%, respectively. This thesis did not attempt to detect severe

cognitive symptoms in patients with depression because only a small number of depression patients (N=4) presented severe cognitive symptoms in our dataset. The classification results of BACS subscales detailed in Table 5.5 were summarized as follows: our analysis achieved balanced accuracies above 70% for BACS-TMT for schizophrenia, BACS-VM, BACS-TMT, and BACS-SC for depression, and BACS-SC for all three groups of participants.

5.3.3 General Psychiatric Symptoms

The BPRS instrument was designed to address general psychiatric symptoms common to schizophrenia and other psychotic disorders. Besides negative symptoms, it also includes hallucinations, delusions, disorganization, hostility, anxiety, and depression. The prediction results of BPRS-Total and the 4-factor scores of BPRS were presented in Table 5.6. For general psychiatric symptoms beyond borderline (BPRS-Total ≥ 24), the BAC is 75.1% when tested on all three groups of subjects combined. For mild and more severe symptoms (BPRS-Total > 32), the BAC for all samples is 73.6% (see Table 5.6). However, the prediction results of BPRS-Total on the schizophrenia group and depression group are not promising (BAC<70%), which may be attributed to the lack of patients with low BPRS-Total. For the factor scores of BPRS, as shown in in Table 5.6, the following results are obtained: for BPRS-NEG, the BAC is 77.7% on all three types of subjects combined and 69.8% for schizophrenia only; for BPRS-POS (BAC=78.9%) and BPRS-RES (BAC=74.1%) of patients with depression, where both scores do not have significant differences from the healthy group (Table 3.1); however, the prediction results of BPRS-AFF and BPRS-NEG for patients with depression are relatively poor. In addition, the prediction results of PANSS, whose BAC is above 70%, are shown in Table 5.7.

5.4 Cross-Site Validation

In our experiment, 54 patients with schizophrenia and 26 health controls were recruited in the first study (Study-A) between 2014 to 2015, and 49 patients with schizophrenia, 50 patients with depression, and 49 healthy controls were recruited in the second study (Study-B) between 2017 and 2018, as shown in Figure 3.1.

TABLE 5.6: Results for automated prediction of the severity of the general psychiatric symptoms assessed by BPRS.

Samples	Scale	THR	Feature	CM		SEN	SPE	F1	MCC	AUPRC	ACC	BAC	MB
				Predicted									
				L	H								
S	BPRS-Total	32.00	F	L 12	12	0.750	0.500	0.609	0.256	0.631	0.614	0.625	0.545
				H 5	15								
	BPRS-AFF	8.39	V	L 36	13	0.592	0.735	0.662	0.330	0.627	0.663	0.663	0.500
				H 20	29								
	BPRS-POS	6.90	V	L 31	17	0.680	0.646	0.663	0.326	0.630	0.663	0.663	0.510
				H 16	34								
	BPRS-NEG	7.02	N	L 30	21	0.809	0.588	0.691	0.405	0.701	0.694	0.698	0.520
				H 9	38								
	BPRS-RES	4.74	F	L 16	8	0.650	0.667	0.660	0.316	0.622	0.659	0.658	0.545
				H 7	13								
D	BPRS-Total	32.00	V	L 14	8	0.692	0.636	0.667	0.329	0.621	0.667	0.664	0.542
				H 8	18								
	BPRS-AFF	12.37	V	L 15	8	0.600	0.652	0.625	0.252	0.578	0.625	0.626	0.521
				H 10	15								
	BPRS-POS	4.04	VN	L 26	2	0.650	0.929	0.807	0.615	0.751	0.813	0.789	0.583
				H 7	13								
	BPRS-NEG	7.21	B	L 17	6	0.421	0.739	0.584	0.169	0.505	0.595	0.580	0.548
				H 11	8								
	BPRS-RES	3.87	N	L 22	5	0.667	0.815	0.748	0.488	0.735	0.750	0.741	0.563
				H 7	14								
DSH	BPRS-Total	24.00	N	L 61	14	0.688	0.813	0.738	0.477	0.758	0.731	0.751	0.653
				H 44	97								
	BPRS-Total	32.00	VN	L 88	50	0.833	0.638	0.714	0.453	0.772	0.708	0.736	0.639
				H 13	65								
	BPRS-AFF	7.47	VN	L 67	43	0.745	0.609	0.675	0.357	0.674	0.676	0.677	0.509
				H 27	79								
	BPRS-POS	4.63	VN	L 77	40	0.737	0.658	0.695	0.395	0.735	0.694	0.698	0.542
				H 26	73								
	BPRS-NEG	6.04	VNF	L 76	30	0.838	0.717	0.779	0.560	0.780	0.780	0.777	0.525
				H 19	98								
	BPRS-RES	3.87	N	L 82	31	0.544	0.726	0.636	0.274	0.670	0.639	0.635	0.523
				H 47	56								

Note: Each assessment score is divided into Class High (H) and Class Low (L) by a cut-off threshold (THR). The THRs of the four domain scores of BPRS are set as their median. The BPRS-Total THR for normal and borderline illness is set to 24. Besides, the BPRS-Total THR for borderline and mild illness is set to 32. Best prediction results for verbal (V), non-verbal (N), facial expression(F), and body movement (B) feature sets are presented. RS=Restricted Speech; PQ=Poor Quality of Speech; AB=Affective Blunting; AM=Amotivation; CM=Confusion Matrix; SEN=Sensitivity; SPE=Specificity; AUPRC=Area Under Precision-Recall Curve; ACC=Accuracy; BAC=Balanced Accuracy; MB=Majority Baseline.

TABLE 5.7: Results for predicting the PANSS scale for schizophrenia (S), depression (D), and healthy controls (H).

Sample	Score	THR	Feature	CM		SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
				Predicted									
				L	H								
S	PANSS-POS	7.37	N	L 16	8	0.542	0.667	0.603	0.210	0.573	0.604	0.604	0.500
				H 11	13								
	PANSS-NEG	8.45	V	L 17	6	0.760	0.739	0.750	0.499	0.687	0.750	0.750	0.521
				H 6	19								
	PANSS-COG	5.19	B	L 14	9	0.773	0.609	0.687	0.386	0.637	0.689	0.691	0.511
				H 5	17								
	PANSS-DEP	5.55	F	L 15	7	0.591	0.682	0.636	0.274	0.591	0.636	0.636	0.500
				H 9	13								
	PANSS-HOS	2.86	V	L 13	12	0.783	0.520	0.640	0.312	0.642	0.646	0.651	0.521
				H 5	18								
	PANSS-DE	4.33	V	L 20	3	0.800	0.870	0.833	0.670	0.831	0.833	0.835	0.521
				H 5	20								
	PANSS-SA	4.86	V	L 20	6	0.545	0.769	0.662	0.324	0.647	0.667	0.657	0.542
				H 10	12								
	PANSS-FSNS	14.50	V	L 25	9	0.929	0.735	0.801	0.606	0.874	0.792	0.832	0.708
				H 1	13								
	PANSS-Total	52.00	VN	L 14	6	0.714	0.700	0.710	0.410	0.688	0.708	0.707	0.583
				H 8	20								
D	PANSS-POS	3.90	B	L 19	4	0.632	0.826	0.735	0.469	0.766	0.738	0.729	0.548
				H 7	12								
	PANSS-NEG	7.96	V	L 17	8	0.565	0.680	0.624	0.247	0.598	0.625	0.623	0.521
				H 10	13								
	PANSS-COG	4.50	VNF	L 22	4	0.542	0.846	0.692	0.409	0.653	0.700	0.694	0.520
				H 11	13								
	PANSS-DEP	8.97	VN	L 9	15	0.750	0.375	0.547	0.135	0.494	0.563	0.563	0.500
				H 6	18								
	PANSS-HOS	3.24	V	L 19	7	0.636	0.731	0.687	0.369	0.642	0.688	0.684	0.542
				H 8	14								
	PANSS-DE	3.80	F	L 14	11	0.824	0.560	0.667	0.384	0.664	0.667	0.692	0.595
				H 3	14								
	PANSS-SA	4.95	B	L 17	5	0.750	0.773	0.762	0.523	0.724	0.762	0.761	0.524
				H 5	15								
	PANSS-FSNS	9.50	VNF	L 6	2	0.881	0.750	0.868	0.558	0.827	0.860	0.815	0.840
				H 5	37								
	PANSS-FSNS	14.50	N	L 20	16	0.833	0.556	0.649	0.338	0.711	0.625	0.694	0.750
				H 2	10								
	PANSS-Total	52.00	V	L 19	7	0.591	0.731	0.665	0.325	0.592	0.667	0.661	0.542
				H 9	13								
DSH	PANSS-POS	3.99	F	L 52	21	0.517	0.712	0.623	0.234	0.621	0.626	0.615	0.557
				H 28	30								
	PANSS-NEG	7.16	VN	L 49	21	0.775	0.700	0.737	0.476	0.734	0.738	0.737	0.504
				H 16	55								
	PANSS-COG	4.48	V	L 47	22	0.625	0.681	0.652	0.306	0.682	0.652	0.653	0.511
				H 27	45								
	PANSS-DEP	5.78	VNFB	L 57	17	0.595	0.770	0.680	0.371	0.689	0.682	0.682	0.500
				H 30	44								

TABLE 5.7: Results for predicting the PANSS scale for schizophrenia (S), depression (D), and healthy controls (H).

Sample	Score	THR	Feature	CM		SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
				Predicted									
				L	H								
DSH	PANSS-HOS	3.08	B	L 49	27	0.571	0.645	0.615	0.215	0.554	0.614	0.608	0.576
				H 24	32								
	PANSS-DE	3.51	VNFB	L 56	20	0.750	0.737	0.743	0.487	0.795	0.743	0.743	0.514
				H 18	54								
	PANSS-SA	4.03	VNFB	L 59	16	0.603	0.787	0.693	0.397	0.697	0.696	0.695	0.507
				H 29	44								
	PANSS-FSNS	9.50	VNF	L 43	9	0.729	0.827	0.769	0.532	0.808	0.764	0.778	0.649
				H 26	70								
	PANSS-FSNS	14.50	VN	L 97	17	0.630	0.851	0.816	0.442	0.850	0.809	0.740	0.809
				H 10	17								
	PANSS-Total	38.00	VN	L 31	8	0.706	0.795	0.744	0.452	0.810	0.730	0.750	0.723
				H 30	72								
	PANSS-Total	52.00	V	L 65	26	0.780	0.714	0.743	0.474	0.758	0.738	0.747	0.645
				H 11	39								

Note: Each assessment score is divided into class *High* (H) and class *Low* (L) by a cut-off threshold (THR). The THR of PANSS-Total and PANSS-FSNS for normal and borderline illness is set to 38 and 9.5, respectively [240, 242]. The PANSS-Total THR for borderline and mild illness is set to 52 and 14.5, respectively [240, 242]. The THRs of other PANSS-related scores are set as the median on the training set. Best prediction results for verbal (V), non-verbal (N), facial expression (F), and body movement (B) feature sets are presented. Abbreviations: POS=positive; NEG=Negative; COG=Cognitive; DEP=Depression/Anxiety; DE=Diminished expression; SA=Social amotivation; FSNS=factor score of negative symptoms; CM=Confusion Matrix; SEN=Sensitivity; SPE=Specificity; F1=F1-score; MCC=Matthews Correlation Coefficient; AUPRC=Area under precision-recall curve; ACC=Accuracy; BAC=Balanced Accuracy; MB=Majority Baseline.

Study-A contains three sessions to initially verify the effectiveness of Cognitive Remediation Therapy (CRT). The three sessions were held at week 0 (Study-A1), week 2 (Study-A2), and week 12 (Study-A3), where about half of the patients with schizophrenia went through CRT after the first clinical assessment at week 0. NSA-16 evaluated the negative symptoms of all participants over three sessions of Study-A. Additionally, due to the participant withdrawal and the malfunction of the recording, only 54 patients with schizophrenia and 24 healthy controls' audio and Kinect recordings were successfully recorded in the Study-A2, and 49 patients with schizophrenia and 22 healthy controls' audio and Kinect recordings were successfully recorded in the Study-A3.

The classification results (schizophrenia group vs. healthy group) and prediction results (class *Low* vs. class *High* of the negative symptom severity) were validated through two validation tasks. The first task is to validate our models' stability across multiple time points, which was trained on the data collected in Study-A1 and Study-B and validated on Study-A2 and Study-A3. The second task is to verify

our algorithm is stable across two data sets (Study-A1 and Study-B). The cross-site validation results for classification and prediction are presented in Table 5.8 and Table 5.9, respectively. In particular, for the first task, eight feature sets weer extracted (verbal: LIWC, Diction, LDA, and Doc2Vec feature sets; non-verbal: Conversational, openSMILE and DisVoice feature sets; and one body movement feature set) from the recordings of Study-A2 and Study-A3, before utilizing the model training on the Study-A1 and Study-B to test Study-A2 and Study-A3 with leave-one-subject-out cross-validation (LOO-CV). For instance, when leveraging our model to predict whether a person in Study-A2/Study-A3 is a patient with schizophrenia or a healthy person, his features extracted from Study-A1 were removed in the training phase. In the second validation task, the feature sets extracted from Study-A1 (Study-B) were used to train the classifier and predict the samples in Study-B (Study-A1).

TABLE 5.8: Results for classification of schizophrenia patients and healthy controls.

Task	Training Session	Testing Session	CM			SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
				Predicted H	Predicted S								
S vs. H	Study A1 and Study B	Study A2	**H**	21	3	0.875	0.852	0.862	0.694	0.950	0.859	0.863	0.692
			S	8	46								
	Study A1 and Study B	Study A3	**H**	19	3	0.864	0.796	0.822	0.620	0.919	0.817	0.830	0.690
			S	10	39								
	Study A1	Study B	**H**	34	15	0.694	0.735	0.714	0.429	0.750	0.714	0.714	0.500
			S	13	36								
	Study B	Study A1	**H**	18	8	0.692	0.889	0.823	0.594	0.834	0.825	0.791	0.675
			S	6	48								

Abbreviations: CM=Confusion Matrix; SEN=Sensitivity; SPE=Specificity; F1=F1-score; MCC=Matthews Correlation Coef
BAC=Balanced Accuracy; MB=Majority Baseline.

TABLE 5.9: Results for predicting the severity of negative symptoms evaluated on depression and schizophrenia groups (DS) and on depression, schizophrenia and healthy groups (DSH).

Samples	Scores	Training Session	Testing Session	THR	Feature	CM			SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
							Predicted L	Predicted H								
DS	NSA-Total	Study A1 and Study B	Study A2	41.0	VNB	**L**	18	7	0.793	0.720	0.759	0.515	0.779	0.759	0.757	0.537
						H	6	23								
		Study A1 and Study B	Study A3	41.0	VNB	**L**	19	6	0.792	0.760	0.776	0.552	0.779	0.776	0.776	0.510
						H	5	19								
		Study A1	Study B	41.0	VNB	**L**	41	8	0.560	0.837	0.691	0.412	0.689	0.697	0.698	0.505
						H	22	28								
		Study B	Study A1	41.0	VNB	**L**	18	9	0.667	0.667	0.667	0.333	0.747	0.667	0.667	0.500
						H	9	18								
DSH	NSA-Total	Study A1 and Study B	Study A2	38.0	VNB	**L**	30	11	0.757	0.732	0.744	0.488	0.768	0.744	0.744	0.526
						H	9	28								
		Study A1 and Study B	Study A3	38.0	VNB	**L**	38	3	0.600	0.927	0.781	0.570	0.814	0.789	0.763	0.577
						H	12	18								
		Study A1	Study B	38.0	VNB	**L**	45	16	0.644	0.738	0.685	0.376	0.667	0.682	0.691	0.588
						H	31	56								
		Study B	Study A1	36.0	VNB	**L**	34	11	0.657	0.756	0.712	0.414	0.713	0.713	0.706	0.563
						H	12	23								

Note: Each assessment score is divided into class *High* (H) and class *Low* (L) by a cut-off threshold (THR). The THR of the NSA-Total
Symptom Assessment; CM=Confusion Matrix; SEN=Sensitivity; SPE=Specificity; F1=F1-score; MCC=Matthews Correlation C
BAC=Balanced Accuracy; MB=Majority Baseline.

To validate our models' stability across multiple time points, a BAC of 86.3% (AUPRC=0.950) was obtained for classifying schizophrenia and healthy control groups of Study-A2 and a BAC of 83.0% (AUPRC=0.919) when tested on Study-A3. Similarly, the prediction results of the severity of the negative symptoms were validated on Study-A2 and Study-A3, which were trained and tested on the samples with and without healthy controls. The model achieved good prediction results in all cases with BACs ranging from 74% to 78% (AUPRC from 0.77 to 0.81). When cross-validate the classification results across two data sets, the BAC was 71.4% (AUPRC=75.0%) when Study-A1 was used to predict Study-B, and when Study-B was used to predict Study-A1, the BAC achieved 79.1% (AUPRC=83.4%). For predicting negative symptom severity, the BACs of including and excluding healthy controls are from 67% to 70% (AUPRC from 0.67 to 0.75).

The above results highlight that the proposed model achieves high and stable performance when being evaluated on multiple time points and between two studies. To investigate the stability of behavioral cues and classification systems across two studies (Study A and Study B), classification (schizophrenia patients vs. healthy controls) and negative symptoms severity prediction (Low vs. High) were validated on Study A and Study B separately and on individual sessions of Study A. Moreover, this might be the first study to investigate how well a machine learning pipeline for digital behavioral phenotyping generalizes across different time points and different studies. Furthermore, the proposed system performs consistently across different time points (the last two sessions from Study A) and for the two independent cohorts (Study A vs. Study B) for recognizing schizophrenia from HC and predicting the severity of negative symptoms. Taken together, these results seem to support our long-term goal of designing low-cost recording technologies for the continuous monitoring of patients.

5.5 Salient Features

Following the method described in Section 3.8, this thesis explored the most salient lexical features to understand the differences in the speech produced by patients with schizophrenia, patients with depression, and healthy controls. The Kruskal-Wallis test was applied to all linguistic features and computed the corresponding

post-corrected p-values, and sorted the top 5 salient categories from low to high in Table 5.10.

TABLE 5.10: Top 5 salient features for video-based modality in paired classification tasks between schizophrenia (S), depression (D), and healthy control (H) groups.

Task	Facial Expressions	Body Movement
$\overline{D}$ vs. H	stuckOutTongue_diff_max^{f}, **Fear_min**^{f*}, cheekRaise_diff_std^{f}, Smile_diff_max^{f*}, MouthOpen_diff_median^{f}, AU06_r_std^{g*}, gaze_angle_y_kurtosis^{g}, **p_3_skew**^{g}, AU12_r_std^{g*}, **p_14_std**^{g}, valence_skew^{h**}, pitch_diff_std^{h}, **arousal_diff_max**^{h}, **arousal_diff_kurtosis**^{h}, **valence_diff_max**^{h}	spine_mid_lin_speed_std^{i}, left_elbow_angle_speed_mean^{i}, right_shoulder_angle_speed_mean^{i}, elbow_right_lin_speed_std^{i}, elbow_left_lin_speed_mean^{i}
$\overline{S}$ vs. H	laughing_diff_mean^{f**}, **LipStretch_diff_max**^{f}, laughing_max^{f**}, **LipStretch_diff_skew**^{f*}, **LipStretch_diff_kurtosis**^{f*}, AU25_c_mean^{g*}, gaze_angle_x_diff_std^{g*}, AU02_c_mean^{g*}, p_scale_skew^{g}, **gaze_angle_x_skew**^{g}, **lambda_kurtosis**^{h}, **yaw_diff_skew**^{h}, valence_skew^{h}, **arousal_diff_max**^{h}, **roll_diff_skew**^{h}	head_lin_speed_mean^{i}, **ankle_left_lin_speed_std**^{i}, left_wrist_angle_speed_mean^{i}, **left_shoulder_angle_acc_mean**^{i}, spine_mid_lin_speed_mean^{i}
$\overline{D}$ vs. S	**EyeClosure_std**^{f*}, **Valence_diff_max**^{f}, **relaxed_diff_max**^{f}, **NoseWrinkle_diff_kurtosis**^{f}, **relaxed_diff_kurtosis**^{f}, **AU01_r_mean**^{g**}, **p_3_skew**^{g}, **gaze_angle_x_diff_mean**^{g*}, p_16_skew^{g}, **gaze_angle_x_diff_std**^{g**}, **pitch_diff_max**^{h}, valence_skew^{h}, **lambda_diff_max**^{h}, lambda_diff_skew^{h}, **yaw_diff_skew**^{h}	left_wrist_angle_speed_mean^{i}, elbow_left_lin_speed_std^{i*}, head_lin_speed_mean^{i}, right_elbow_angle_acc_mean^{i*}, spine_mid_lin_speed_std^{i*}

Feature set: a-LIWC; b-Diction; c-Conversational; d-OpenSmile; e-DisVoice; f-Affectiva; g-OpenFace; h-Opsis
Suffix: min-minimum; max-maximum; sk-skewness; ku-kurtosis; std-standard deviation; de-delta value; linregc-linear regression coefficients; clf-classification value; reg-regression value.
**: p-value ≤ 0.005; *: p-value ≤ 0.05; Kruskal-Wallis tests with FDR correction.
bold feature: the average value of this feature is larger in the class with overline.

By analyzing the facial expression features, we noted that the expression of positive emotions in patients with depression is reduced compared with healthy controls, as reflected by the smaller delta values of Smile, CheeckRaise, and MouthOpen features. Although these features are not statistically different because the p-values of about five thousand features have been corrected, it was found that smile intensity and smile duration are important for detecting depression [234]. Moreover, we also noted that patients with depression had a longer duration of looking down, which caused a smaller kurtosis value on the vertical gaze angle and smaller mean and STD values of EyeClosure compared with healthy controls. Those abnormal eye movements were also observed by Girard and his colleagues, which may indicate fatigue [52].

For patients with schizophrenia, the most interesting pattern is Laughing. Laughter was evaluated and represented through a regression value (range from 0 to 100) measured by Affectiva, and this metric is referred to as 'Laughing'. Meanwhile, the

mean, standard deviation, and maximum values of the delta value of 'Laughing' for schizophrenia patients were smaller than those of healthy participants. These differences reflect patients with schizophrenia expressed less laughter than the control group [55]. Furthermore, patients with schizophrenia had a smaller gaze angle in the horizontal direction (GazeAngle_x) than healthy controls, which might result from the fact that they rarely looked at their surroundings.

Moreover, depression patients presented higher STD values of EyeClosure (15.76, 95% CI - 13.12 to 18.35; $P<0.05$) compared with schizophrenia patients, which indicate that depression may feel more fatigued during the interview [52]. Besides, we noted that patients with depression had a smaller standard deviation of the linear speed of the left elbow (1.15, 95% CI - 0.92 to 1.39; $P<0.005$) and spine (0.49, 95% CI - 0.35 to 0.63; $P<0.005$), which indicates that patients with depression have fewer upper body and arm movements than patients with schizophrenia during the interview.

5.6 IMH dataset vs. DAIC-WOZ

In this section, we compare the prediction results of the severity of depression symptoms between the dataset we collected in IMH and the DAIC-WOZ dataset [158]. The audio and video of the participants' conversations were recorded for these two datasets. Both datasets measured the severity of depressive symptoms using the Patient Health Questionnaire (PHQ) scale. In addition, the DAIC-WOZ dataset does not provide video recording but provides features extracted from the video by OpenFace. There are also some differences between the two data sets: 1) DAIC-WOZ dataset contains interview recordings between participants and avatars, but the recordings in our dataset are face-to-face recordings between participants and psychiatrists; 2) participants in the DAIC-WOZ dataset were recruited from the U.S. Armed Forces and the public and were coded as depression, post-traumatic stress disorder (PTSD), and anxiety, but the participants in the IMH data set were formally diagnosed with depression and schizophrenia using DSM-IV; 3) the participants of the DAIC-WOZ dataset are from the United States, while the population of the IMH dataset is mainly Asia; 4) we assessed the depressive symptoms using PHQ-9, but the experiment for DAIC-WOZ dataset used PHQ-8; 5the transcriptions of DAIC-WOZ dataset were manually transcripted by a senior transcriber,

but we applied an automated ASR toolkit to automatically convert the speech to text; 6) The DAIC-WOZ dataset interviews are structured (all participants were asked the same questions), while interviews in this thesis are semi-structured (open discussion based on a question set).

According to the methods introduced in Section 5.1 and Section 4.1, 4 verbal feature sets (LIWC, Diction, LDA, and Doc2Vec), 2 non-verbal feature sets (DisVoice and OpenSmile), and 1 facial expression feature set (OpenFace) from the DAIC-WOZ dataset were extracted. Next, we implemented the ensemble classification methods with LOO-CV (see Section 3.6) to predict the severity of depressive symptoms measured by the PHQ scale (Low vs. High) for the IMH dataset and DAIC-WOZ dataset. The PHQ scores of each dataset were divided into two balanced categories according to the median value, where the PHQ cut-off score of the IMH dataset and DAIC-WOZ dataset is 10 and 7, respectively.

TABLE 5.11: Results for predicting the severity of depressive symptoms of IMH dataset and DAIC-WOZ dataset.

	Modality	**IMH dataset**			**DAIC-WOZ [158]**		
		F1	**AUPRC**	**BAC**	**F1**	**AUPRC**	**BAC**
Single modality	LIWC	0.600	0.632	0.596	0.608	0.633	0.615
	Diction	0.665	0.674	0.638	0.665	0.649	0.666
	Doc2Vec	0.563	0.595	0.527	0.578	0.607	0.582
	LDA	0.603	0.590	0.553	0.608	0.599	0.611
	DisVoice	0.621	0.628	0.633	0.603	0.615	0.603
	OpenSMILE	0.699	0.669	0.654	0.571	0.604	0.570
	OpenFace	0.692	**0.724**	0.666	0.587	0.556	0.595
Late fused modality	Verbal	0.678	0.654	0.649	0.679	0.677	0.680
	Non-verbal	0.676	0.681	0.697	0.629	0.627	0.631
	Audio	0.696	0.696	0.697	**0.686**	**0.701**	**0.691**
	Audio + Video	**0.726**	0.707	**0.703**	0.679	0.699	0.680

Verbal: LIWC, Diction, LDA, and Doc2Vec.
Non-verbal: DisVoice and OpenSmile.
Audio: Verbal and Non-verbal.
Video: OpenFace.

The prediction results of each single feature set and the fusion results are presented in Table 5.11. Using audio and video feature sets, we observed that the prediction accuracy of predicting PHQ scores on both data sets is about 70%. Moreover, the classification result of using language information on DAIC-WOZ data is better than the IMH dataset. The reason might be the DAIC-WOZ dataset interviews are structured, and a human transcribed the speech. In addition, the performance of using non-verbal and facial expression modalities on the IMH dataset is better

than the DAIC-WOZ dataset. It might be because the participants in the DAIC-WOZ experiment made fewer facial expressions as they are facing an avatar rather than an actual human.

5.7 Our methods vs. state-of-the-art methods

In this section, we implement three state-of-the-art methods (VGG [262], DenseNet [263], and Sch-net [264]) on the dataset that we collected in IMH and compared the classification results between these state-of-the-art methods and our methods. We validate these three methods using 10-fold cross-validation (CV). In each CV loop, 10% samples were left out as the test set, and the rest of 80% and 20% samples were split into training and validation sets. For all three methods, the model was trained on the training set with a large epoch number, and the model with the lowest validation loss will be used to predict the test set. Specifically, the parameters of VGG and DenseNet neural networks were fixed because these two neural networks were pre-trained with the Affwild dataset [265], and then fine tuned the classification layer using appropriate learning rate. For Sch-net, since the pre-trained model was not available, we retrained both Sch-Net and the classification layer on our dataset. In the subsequent paragraphs, we briefly introduce the key information of these three methods and compare the classification results with ours.

The VGG-16 and DenseNet-201 features were used as baseline feature sets for depression detection in the AVEC 2019 challenge [162]. The VGG-16 and DenseNet-201 neural networks are two CNN-based deep representation learning paradigms commonly used in the image processing area. In speech classification tasks, the Mel-Spectrograms of speech instances are fed into the pre-trained VGG and DenseNet neural network, and a set of the resulting activations are extracted as feature vectors. The detailed parameters and methods implemented in our classification tasks can refer to [162]. The feature vectors are fed into the classification layer to achieve the classification task, which is composed of a fully connected neural network with a ReLU activation function and a softmax layer.

Sch-net is a CNN-based deep learning architecture for schizophrenia detection [264]. The spectrogram is given as the input of the Sch-net. The convolutional layer is

used to capture the local features in the spectrogram. Other mechanisms, e.g., average pooling operation, skip connections, and attention module, are also employed in Sch-net [264]. We applied same parameter settings and the neural network structure of Sch-net on our dataset. The learning rate and epoch number were

TABLE 5.12: Performance of the participant group classification using the state-of-the-art methods and proposed methods.

	Feature Set	IMH dataset								
		S vs. H			D vs. H			D vs. S		
		F1	AUPRC	BAC	F1	AUPRC	BAC	F1	AUPRC	BAC
Proposed method	Verbal	0.797	0.849	0.784	0.731	0.799	0.729	0.720	0.788	0.722
	Nonverbal	0.780	0.778	0.771	0.790	0.817	0.792	0.780	0.857	0.784
	Audio	0.821	0.867	0.809	0.798	0.861	0.806	0.817	0.858	0.804
	Audio + Video	0.835	0.874	0.821	0.800	0.865	0.802	0.829	0.893	0.834
	Audio + Video + Movement	**0.836**	**0.890**	**0.823**	**0.838**	**0.879**	**0.823**	**0.858**	**0.905**	**0.847**
State-of-the -art method	VGG [162]	0.622	0.638	0.622	0.647	0.744	0.657	0.690	0.814	0.755
	DenseNet[162]	0.645	0.717	0.653	0.622	0.682	0.629	0.759	0.811	0.753
	Sch-net[264]	0.604	0.607	0.605	0.638	0.724	0.647	0.752	0.829	0.792

We present the classification results of the above mentioned three methods and proposed methods in Table 5.12. We observe that the conventional machine learning methods with feature engineering that we implement in this thesis can achieve significantly better results than the complex deep learning methods. The main reason might be that the features extracted from the Mel spectrogram cannot distinguish between mentally ill and healthy controls in our dataset.

5.8 Conclusion

As described before, our study cooperated with IMH Singapore and collected audio interview recordings from a total of 103 schizophrenia patients, 50 depression patients, and 75 healthy controls. This thesis employed video signal processing, facial emotion/expression recognition, body movement analysis, and machine learning algorithms to automatically analyze the video interview recordings of schizophrenic patients, depression, and healthy control subjects on two different datasets. Moreover, this thesis investigated the proposed digital phenotype models are consistent and stable across recording sessions at different time points and across different studies and datasets, which constitutes a first small step towards automated longitudinal follow-up of negative (and other) symptoms in psychiatric patients. This approach could further be applied in long-term video monitoring and diagnosis of

mentally ill patients. For a better understanding of the core results of this chapter, we summarized the classification results and prediction results in Table 5.13.

TABLE 5.13: Summarized results for automated classification of schizophrenia, depression, and healthy controls using audiovisual modalities.

Mode	Modality	Classification tasks				Prediction tasks					
						Negative symptoms		Cognitive symptoms		General psychiatric symptoms	
		S vs. H	D vs. H	D vs. S	DS vs. H	Global score	Total score	Normal vs. Mild	Mild vs. Severe	Normal vs. Mild	Mild vs. Severe
Single modality	Affectiva	0.744	0.654	0.671	0.644	0.668	0.614	0.606	0.646	0.549	0.566
	OpenFace	0.744	0.615	0.731	0.707	0.669	**0.797**	0.670	0.698	0.643	0.563
	Opsis	0.589	0.702	0.628	0.642	0.632	0.628	0.514	0.494	0.660	0.523
	Movement	0.652	0.686	0.710	0.608	0.530	0.505	0.651	0.566	0.528	0.549
Fusion of modalities	Facial	0.744	0.663	0.744	0.732	0.664	0.729	0.660	0.649	0.703	0.562
	Speech + Facial	0.821	0.802	0.834	0.788	**0.722**	0.760	0.765	0.725	**0.726**	0.669
	Speech + Facial + Movement	**0.823**	**0.823**	**0.847**	**0.798**	0.716	0.750	**0.789**	**0.735**	0.711	**0.671**

Facial modality: Affectiva, OpenFace, and Opsis.
Speech modality: LIWC, Diction, LDA, Doc2Vec, Conversational, DisVoice, OpenSmile, Affectiva, OpenFace, and Opsis.

Along with the audio analysis described in Chapter 4, the proposed machine learning system achieved a moderate-high accuracy for classifying the total score of negative symptoms (BAC, 76.0%; SEN, 80.2%; SPE, 71.8%), the composite score of cognitive symptoms (BAC, 75.6%; SEN, 80.8%; SPE, 70.5%), and total score of general psychiatric symptoms (BAC, 73.6%; SEN, 83.3%; SPE, 63.8%). Compared with only applying features extracted from audio, the fusion of facial expression features and body movement features can promote the classification accuracy of diagnosis by about 2%-4%. Our results notably demonstrated the success of predicting assessment ratings that are directly or indirectly related to diminished expressions with a moderate-high BAC (>75%), such as restricted speech, affective blunting, and token motor test, while achieving relatively poor results (<65%) on semantic fluency and resistance factor scores, which are not directly related to diminished expression. Furthermore, the proposed system is able to differentiate schizophrenia and depression recordings from healthy control recordings with 82.3% BAC, differentiate between depression and schizophrenia with a BAC of 84.7%, and distinguish the three groups combined (schizophrenia, depression, and healthy controls) with a 3-class classification accuracy of 68.7%.

Besides, many salient visual behavioral cues were observed between patients and controls, as well as between schizophrenia and depression. These observations suggest that our set of audio-video markers is capable of tracking clinically relevant

behaviors and behavioral changes. More research is needed to confirm these findings on multiple independent patient populations, and more extensive longitudinal studies are required to investigate trends in the behavioral changes due to various interventions. These results are promising and present an important step towards our overall goal of creating automated systems to aid clinical diagnosis and understanding of schizophrenia and depression.

Chapter 6

Speech Adaptation Analysis for Mental Disorders

In this chapter, the IAT theory was referenced [64] to quantify the speech adaptation between patients with schizophrenia, patients with depression, and healthy controls with their interlocutors. This chapter build on the methods described in previous chapters and showcase that the automated system harnessing speech adaptation cues can predict the symptom severity of patients with schizophrenia and depression and classify different groups of participants. This chapter first demonstrates how the audio recordings collected in semi-structured clinical interviews were pre-processed. This chapter then introduces the method to extract speech adaptation features from turn-takings. Next, the impact of two hyperparameters (minimum duration and maximum gap) on the differentiation of patients and healthy control and the selection of these two hyperparameters are presented. Then, this thesis presents the performance on predicting NSA-16 and distinguishing patients from healthy controls. Finally, the most predictive speech adaptation features between different groups of participants (depressed, schizophrenia, and healthy controls) are introduced. The pipeline of speech adaptation analysis of mentally ill patients is shown in Figure 6.1.

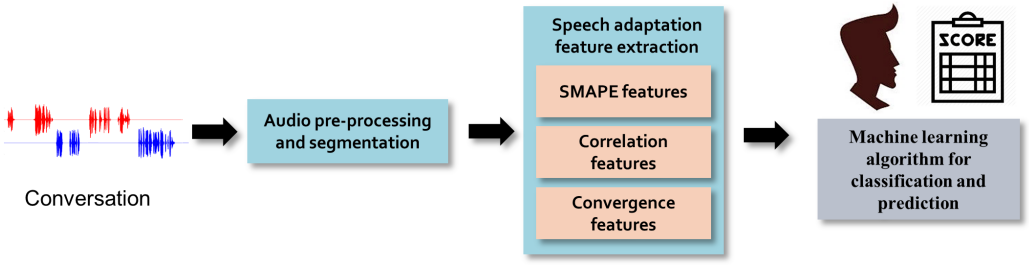


FIGURE 6.1: Diagram of the speech adaptation analysis pipeline.

6.1 Feature Extraction

6.1.1 Audio Pre-processing

Before identifying speaking utterances from the recordings, three pre-processing steps. First, the segments recorded during the installation and removal of the recording equipment were removed. Second, the noise reduction function in Audacity was applied on each channel of the audio recording to reduce the noise by typically -6 dB. The noise statistics were automatically extracted from a manually selected noisy segment and the algorithm is described in the Audacity wiki. Thirdly, the volume of all recordings was automatically normalized into -26 dB loudness level using the EBU R128 loudness normalization procedure [266] using FFmpeg-normalize package[1]. This volume normalization is applied on two-channel speech signals to remove the effect of data volume on the results.

Since both participant's and psychologists' speeches were recorded in a meeting room simultaneously, both channels are mixed with the voices of two people. In order to automatically extract the speech of both participants and psychologists, speaker diarization techniques were then applied, which have been described in sec:speaker-diarization.

6.1.2 Segmentation

After the speaker diarization, an energy-based voice activity detection (VAD) library [2] was leveraged to further identify the speech and non-speech parts for both

[1]https://github.com/slhck/ffmpeg-normalize
[2]https://github.com/jiaaro/pydub

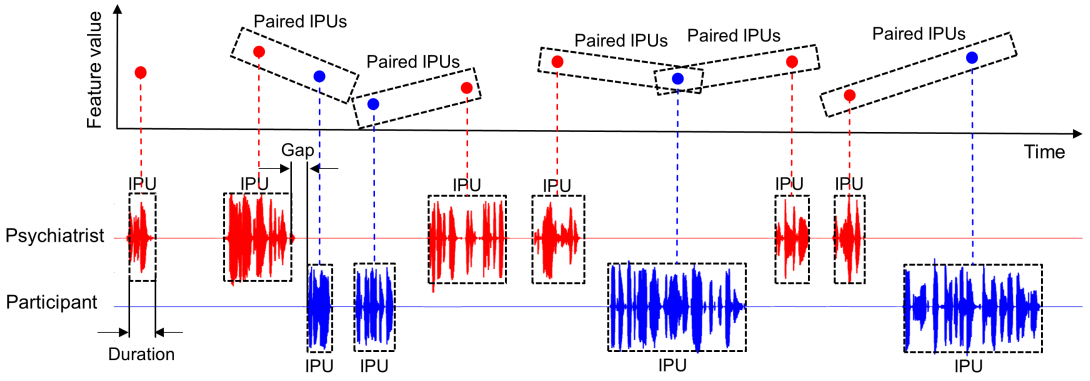


FIGURE 6.2: Schematic of IPU and paired IPUs.

participant and psychiatrist channels. The upper bound of the silent is set to -50 dBFS and the minimum length of the silent segment is set to 0.5.

Similar to [267], the speech adaptation in the conversation of two people, Speaker A and Speaker B, is measured through a turn-by-turn manner. The schematic of inter-pausal unit (IPU) and paired IPUs are shown in Figure 6.2. One turn refers to the period of time when one speaker starts speaking until another person responds, which is composed of one or multiple IPUs. Moreover, one paired IPUs is defined as the two most adjacent IPUs (one paired IPUs), which comprise the last IPU in one speaker's turn and the first IPU in the consequent turn of another speaker.

In addition, two hyperparameters that may affect the classification results were defined in pre-processing step: the minimum duration of IPUs (Min_Dur) and the maximum gap between two adjacent IPUs (Max_Gap). The gap and duration of IPU have illustrated in Figure 6.2. The Min_Dur controls the length of the IPUs before constructing the paired IPUs. The IPUs whose duration is smaller than this threshold will be filtered out. The Max_Gap determines the maximum interval of all paired IPUs. How these two hyperparameters influence the distinction between patients and healthy controls is discussed in Section 6.1.6.

6.1.3 Acoustic/Prosodic Features

The openSMILE toolkit [268] is a modular and adjustable collection of acoustic features useful for signal processing and machine learning applications. In

TABLE 6.1: Summary of LLDs used in eGeMAPS parameter settings.

LLDs Name (Abbreviation)	Description
Frequency related parameters	
Pitch (F0)	logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
Jitter	deviations in individual consecutive F0 period lengths
1-3 Formant Frequency (FormantFreq)	centre frequency of first, second, and third formant
1-3 Formant Bandwidth (FormantBand)	bandwidth of first, second, and third formant
Energy related parameters	
Shimmer	difference of the peak amplitudes of consecutive F0 periods
Loudness	estimate of perceived signal intensity from an auditory spectrum
Equivalent Sound Level (Leq)	the same total sound energy being produced over a given period
Harmonics-to-Noise Ratio (HNR)	relation of energy in harmonic components to energy in noise-like components
Spectral parameters	
Alpha Ratio (AlphaR)	ratio of the summed energy from 50–1000 Hz and 1–5 kHz
Hammarberg Index (HamIndex)	ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region
Spectral Slope 0–500 and 500–1500 Hz	linear regression slope of the logarithmic power spectrum within the two given bands
Relative 1-3 Formant Energy (RFN)	ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F0
Harmonic Difference H1–H2 (HD1-2)	ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2)
Harmonic Difference H1–A3 (HD1-3)	ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3)
MFCC 1–4	Mel-Frequency Cepstral Coefficients 1–4
Spectral flux (SpecFlux)	difference of the spectra of two consecutive frames
Temporal parameters	
Loudness Peaks Rate (LoudnessPeak)	the number of loudness peaks per second
Voice and Unvoiced Segments Length (VSL and UVSL)	the mean length and the standard deviation of continuously voiced regions (F0 > 0) and unvoiced length (F0 = 0)
VoicedUnvoiced Segments (UVS)	the number of continuous voicedunvoiced regions per second

this work, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) configuration was leveraged to extract the acoustic/prosodic features from IPUs, as eGeMAPS is consistently known to increase the accuracy of automatic affect recognition tasks [269]. It implements prosodic, excitation, vocal tract, and spectral descriptors, including 8 Frequency related parameters, 4 Energy/Amplitude related parameters, 14 Spectral parameters, and 3 temporal parameters (shown in Table 6.1). Moreover, arithmetic mean and variation are applied to the above descriptors. The additional 8 functionals are employed to loudness and pitch: 20th, 50th, and 80th percentile (pct20, pct50, and pct80), the range of percentile (e.g., pct0-20 and pct20-80), and the mean and standard deviation (std) of the signal parts of rising slope (RSlope) and falling slope (FSlope). The above measurements result in a total of 88 acoustic features for each IPU. Finally, the acoustic features of all IPUs were standardized using the z-score on the two channels.

6.1.4 Speech Adaptation Modeling

After extracting acoustic features from paired IPUs of Speaker A and Speaker B, speech adaptation features in psychiatrist-participant dialogues were computed.

In order to measure all of the acoustic characteristics between the psychiatrist and the participants, the speech adaptation patterns were evaluated between every pairwise combination of acoustic characteristics. In the following, three feature representation methods are proposed for capturing the coordinated patterns of the non-verbal features between psychiatrists and participants.

First, the degree of absolute similarity between the acoustic features of interlocutors is calculated utilizing the Symmetric Mean Absolute Percentage Error (SMAPE), which is given by:

$$SMAPE(A, B) = \frac{1}{N} * \sum_{t=1}^{N} \frac{|A_t - B_t|}{|A_t| + |B_t|}, \tag{6.1}$$

where N denotes the number of paired IPUs, A_t the value of acoustic features of speaker A at the t-th paired IPUs, and B_t the value of acoustic feature of speaker B at the t-th paired IPUs. The SMAPE value subtracted by 1 (1- SMAPE) was used to indicate the degree of similarity of the non-verbal patterns between psychiatrists and participants so that the value close to one indicates a higher degree of proximity between conversation partners.

Reciprocity between the acoustic features of interlocutors indicates a change in the nonverbal behavior of one speaker that matches a similar change in the comparable function value of another speaker rather than to match it. This relative similarity is quantified utilizing Pearson correlation via:

$$r(A, B) = \frac{\sum_{t=1}^{n}(A_t - \bar{A})(B_t - \bar{B})}{\sqrt{\sum_{t=1}^{n}(A_t - \bar{A})^2}\sqrt{\sum_{t=1}^{n}(B_t - \bar{B})^2}}, \tag{6.2}$$

where $\bar{A}$ and $\bar{B}$ are the average values of speaker A and speaker B, respectively. The value of Reciprocity is in the range of $[-1, 1]$. A high positive value would be characterized by high levels of reciprocity, and vice versa.

Convergence is defined as one actor's behavior becomes more like another over time, which usually is a sign of agreement and positive evaluation towards the interlocutor. The degree of convergence is quantified by the proportion of how many times the difference between the voice features of the two speakers is smaller than that in the previous turn, which is given by:

$$Convergence(A, B) = \frac{1}{N} * \sum_{t=1}^{N} I(|A_t - B_t| > |A_{t-1} - B_{t-1}|), \tag{6.3}$$

where the formula above provides a result between 0 and 1, and a larger value represents a higher level of convergence.

As mentioned before, these three forms of speech adaptation were evaluated across all the combinations of 88 acoustic features. Therefore, each of the three modalities contains 7744 features. These three groups of features are referred to as SMAPE, Correlation, and Convergence modalities.

6.1.5 Classification Method

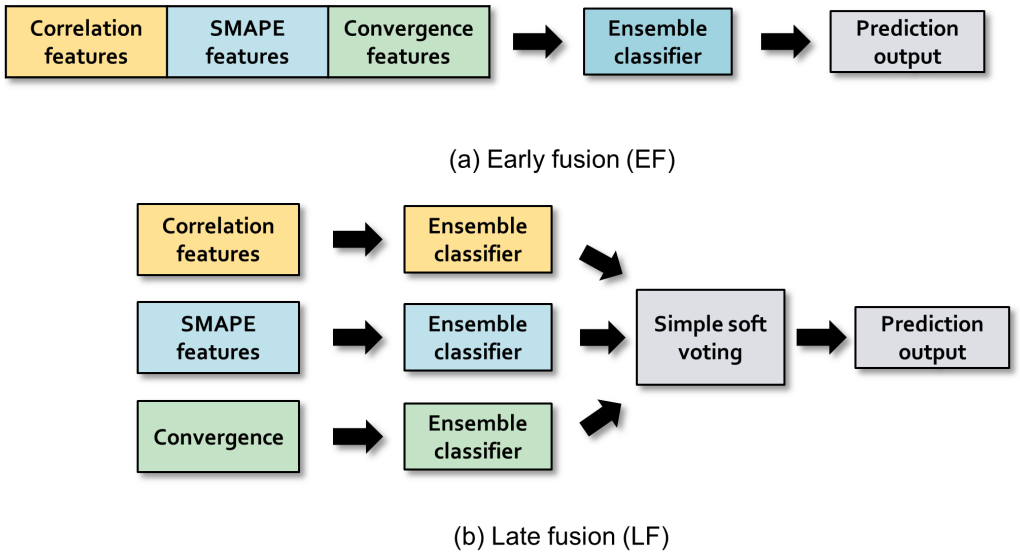


FIGURE 6.3: This study used two multi-modality fusion strategies: (a) Early fusion, (b) Late fusion.

Based on the ensemble classifier described in Section 3.6, this thesis leveraged three speech adaptation modalities (SMAPE, Correlation, and Convergence) with late fusion to classify the three different participant groups (classification tasks) and predict the clinical assessment scales (prediction tasks). Similarly, the ensemble classifier consists of 5 different classifiers: Support Vector Machine, Logistic Regression, Gradient Boosting, AdaBoost, and Random Forest, and those classifiers

were assessed by leave-one-out cross-validation (LOOCV). Specifically, both early and late fusion approaches were considered to integrate SMAPE, Correlation, and Convergence modalities (see Figure 6.3). Early fusion concatenates modalities into a common modality, which is also known as feature level fusion, while late fusion averages the probability outputs of modalities at the decision level.

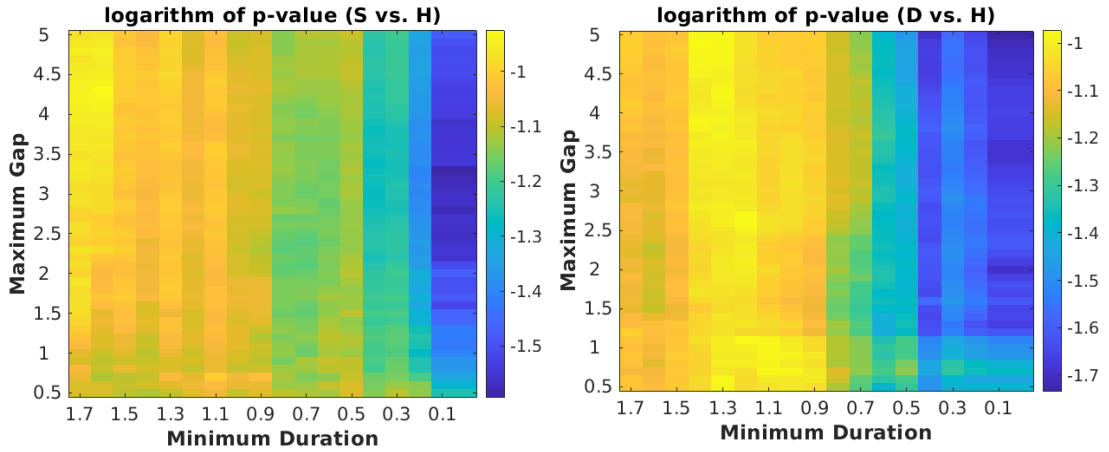


FIGURE 6.4: The top 10th percentile logarithmic p-value of reciprocity features between patients and healthy controls.

6.1.6 Hyperparameter Selection

The conversational timing and the speech duration are the two social elements in human-human conversation [270]. Our work analyzes how these two factors affect the difference in speech adaptation between patients and healthy controls. As described in Section 6.1.2, Min_Dur and Max_Gap are applied to filter out more relatively short voices and control how close the two sentences are. Intuitively, there will not be speech adaptation in dialogue if the pause time is too long or the duration is too short. To indicate how these two hyperparameters affect our classification and prediction results, the Kruskal-Wallis test was applied to calculate the statistical differences for Reciprocity features between patient groups and healthy groups under different parameter settings.

Specifically, the Kruskal-Wallis p-value was computed for each feature in the range of Min_Dur from 0 to 1.7 and Max_Gap from 0.5 to 5 with an interval of 0.1. The top 10th percentile logarithmic p-value of reciprocity features between patients and healthy controls are presented in the color maps in Figure 6.4. It shows that features calculated under different Max_Gap values seem to have little effect in

TABLE 6.2: Results for automated classification of schizophrenia (S), depression (D), and healthy controls (H) using speech adaptation feature sets.

Task	Modality	CM			SEN	SPE	F1	AU-PRC	ACC	BAC	MB
			Predicted								
			H	S							
S vs. H	**Correlation**	**H**	**46**	**24**	**0.657**	**0.827**	**0.754**	**0.780**	**75.6%**	**74.2%**	**0.583**
		S	**17**	**81**							
	SMAPE	**H**	45	25	0.643	0.653	0.651	0.675	64.9%	64.8%	0.583
		S	34	64							
	Convergence	**H**	41	29	0.586	0.541	0.563	0.578	56.0%	56.3%	0.583
		S	45	53							
	All (early fusion)	**H**	46	24	0.657	0.776	0.726	0.735	72.6%	71.6%	0.583
		S	22	76							
	All (late fusion)	**H**	49	21	0.700	0.755	0.733	0.774	73.2%	72.8%	0.583
		S	24	74							
			D	**H**	**SEN**	**SPE**	**F1**	**AU-PRC**	**ACC**	**BAC**	**MB**
D vs. H	Correlation	**D**	31	17	0.646	0.657	0.655	0.685	65.3%	65.1%	0.593
		H	24	46							
	SMAPE	**D**	30	18	0.625	0.629	0.630	0.653	62.7%	62.7%	0.593
		H	26	44							
	Convergence	**D**	24	24	0.500	0.486	0.496	0.499	49.2%	49.3%	0.593
		H	36	34							
	All (early fusion)	**D**	31	17	0.646	0.671	0.663	0.669	66.1%	65.9%	0.593
		H	23	47							
	All (late fusion)	**D**	**33**	**15**	**0.688**	**0.657**	**0.672**	**0.684**	**66.9%**	**67.2%**	**0.593**
		H	**24**	**46**							
			D	**S**	**SEN**	**SPE**	**F1**	**AU-PRC**	**ACC**	**BAC**	**MB**
D vs. S	Correlation	**D**	35	13	0.684	0.729	0.707	0.784	69.9%	70.6%	0.671
		S	31	67							
	SMAPE	**D**	38	10	0.684	0.792	0.728	0.786	71.9%	73.8%	0.671
		S	31	67							
	Convergence	**D**	26	22	0.551	0.542	0.562	0.616	54.8%	54.6%	0.671
		S	44	54							
	All (early fusion)	**D**	35	13	0.684	0.729	0.707	0.784	69.9%	70.6%	0.671
		S	31	67							
	All (late fusion)	**D**	**34**	**14**	**0.776**	**0.708**	**0.758**	**0.804**	**75.3%**	**74.2%**	**0.671**
		S	**22**	**76**							

distinguishing patients from healthy controls. However, a larger Min_Dur value seems to filter out too much content, so using a smaller value may capture more differences between the patient and healthy control groups. With the above, we set the Min_Dur value to 0 and Max_Gap to 3 in the following classification and prediction tasks.

6.2 Classification of Participants

According to the result shown in Table 6.2, we note that it is feasible to use speech adaptation patterns to differentiate depression, schizophrenia, and control groups

for a diagnosis purpose. For classification of participant groups, applying late fusion of speech adaptive feature sets obtains higher balance classification accuracies than the early fusion: schizophrenia vs. healthy (BAC=72.8%), depression vs. healthy (BAC=67.2%), depression vs. schizophrenia (BAC=74.2%). When distinguishing between patients with schizophrenia and healthy controls, the best prediction result was obtained by using the Correlation modality alone, which indicates that there is a significant difference in the correlation between the non-verbal features in the dialogue. However, the Convergence modality has poor classification accuracy across all classification tasks.

TABLE 6.3: Results for automated classification of schizophrenia (S), depression (D), and healthy controls (H) using speech adaptation feature sets.

Task	**Modality**	**CM**			**SEN**	**SPE**	**F1**	**AU-PRC**	**ACC**	**BAC**	**MB**
			Predicted								
			H	**S**							
S vs. H	VN	**H**	55	15	0.786	0.837	0.816	0.879	81.5%	81.1%	0.583
		S	16	82							
	VNA	**H**	59	11	0.843	0.847	0.846	0.885	84.5%	84.5%	0.583
		S	15	83							
	VNFB	**H**	59	22	0.728	0.907	0.824	0.899	82.6%	81.8%	0.545
		S	9	88							
	VNFBA	**H**	**55**	**20**	**0.733**	**0.971**	**0.867**	**0.911**	**87.1%**	**85.2%**	**0.579**
		S	**3**	**100**							
			D	**H**	**SEN**	**SPE**	**F1**	**AU-PRC**	**ACC**	**BAC**	**MB**
D vs. H	VN	**D**	32	16	0.667	0.914	0.809	0.842	81.4%	79.0%	0.593
		H	6	64							
	VNA	**D**	**34**	**14**	**0.708**	**0.900**	**0.819**	**0.843**	**82.2%**	**80.4%**	**0.593**
		H	**7**	**63**							
	VNFB	**D**	38	12	0.760	0.800	0.785	0.864	78.4%	78.0%	0.600
		H	15	60							
	VNFBA	**D**	38	12	0.760	0.813	0.793	0.854	79.2%	78.7%	0.600
		H	14	61							
			D	**S**	**SEN**	**SPE**	**F1**	**AU-PRC**	**ACC**	**BAC**	**MB**
D vs. S	VN	**D**	42	6	0.724	0.875	0.781	0.849	77.4%	80.0%	0.671
		S	27	71							
	VNA	**D**	38	10	0.857	0.792	0.837	0.872	83.6%	82.4%	0.671
		S	14	84							
	VNFB	**D**	42	8	0.806	0.840	0.821	0.893	81.7%	82.3%	0.673
		S	20	83							
	VNFBA	**D**	**43**	**7**	**0.816**	**0.860**	**0.834**	**0.902**	**83.0%**	**83.8%**	**0.673**
		S	**19**	**84**							

In an attempt to test whether the proposed speech adaptation features can improve the classification accuracy, the probability outputs of speech adaptation feature sets with audio-based (verbal and non-verbal) and the outputs of audio-visual (verbal, non-verbal, facial expression, and body movement) feature sets were late fused. In summary, merging speech adaptation modalities with the primary modalities improves the classification performance from 1.4% to 3.4%. The BAC reached 85.2% for schizophrenia vs. healthy, 80.4% for depression vs. healthy, and 83.8% for depression vs. schizophrenia (see Table 6.3). These results demonstrate that there are differences in the interactive vocal characteristics between every two groups, and combining the speech adaptation features with verbal and non-verbal features is able to improve the group-level identification. To gain insight into what distinguishes the different participants and interpret our findings, we further analyzed the significant features in Section 6.4.

6.3 Prediction of Symptom Severity

Similarly, the prediction of clinical assessment scores was also conducted using speech adaptation features as input and clinical scores as targets. As described in section 3.5, The clinical scores were conducted into 2 classes (Class High and Class Low) based on different cut-off scores. Then, speech adaptation features were employed as input, and the clinical rating (low or high) was treated as the label. In addition to using speech adaptation features alone for prediction, this thesis also compares the prediction results after integrating the speech adaptation features with verbal, non-verbal, facial, and body movement feature sets introduced in chapter 4 and chapter 5. In Table 6.4, The prediction results were presented for the overall severity of negative, cognitive, and general psychiatric symptoms assessed by NSA-16, BACS, and BPRS, respectively.

From the results, speech adaptation features seem to have a very small promotion effect in predicting negative symptoms. After fusing the outputs of speech adaptation feature sets and the audio-visual feature sets, the classification accuracy for distinguishing negative emotions is 75.9%, which is 0.9% compared with the classification accuracy of using VNFB feature sets. For distinguishing the severity of cognitive symptoms, using these features is not able to improve the classification performance of the previous audio-visual system. Instead, after integrating speech

TABLE 6.4: Results for automated prediction of the severity of the negative, cognitive, and general psychiatrist symptoms using verbal (V), nonverbal (N), facial (F), body movement (B), and speech adaptation (A) cues.

Scale	THR	Modality	CM			SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
				Predicted									
				H	L								
NSA-Total	38	A	L	71	42	0.628	0.680	0.653	0.308	0.666	65.3%	65.4%	0.523
			H	33	70								
		VN	L	87	26	0.770	0.738	0.755	0.508	0.784	75.5%	75.4%	0.523
			H	27	76								
		VNA	L	93	20	0.823	0.670	0.748	0.500	0.782	75.0%	74.6%	0.523
			H	34	69								
		VNFB	L	90	29	0.756	0.743	0.750	0.499	0.792	75.0%	75.0%	0.522
			H	28	81								
		VNFBA	L	89	30	0.748	0.771	0.759	0.518	0.784	75.9%	75.9%	0.522
			H	25	84								
BACS-Composite	-1	A	L	58	66	0.739	0.468	0.580	0.211	0.601	58.3%	60.3%	0.574
			H	24	60								
		VN	L	90	34	0.772	0.726	0.747	0.492	0.821	74.5%	74.9%	0.574
			H	21	66								
		VNA	L	93	31	0.685	0.750	0.723	0.434	0.775	72.2%	71.7%	0.574
			H	29	34								
		VNFB	L	108	24	0.760	0.818	0.794	0.578	0.821	79.4%	78.9%	0.579
			H	23	31								
		VNFBA	L	92	40	0.771	0.697	0.730	0.462	0.783	72.8%	73.4%	0.579
			H	22	24								
	-2	A	L	92	40	0.568	0.651	0.668	0.181	0.743	63.4%	61.0%	0.796
			H	13	31								
		VN	L	112	60	0.705	0.808	0.800	0.452	0.853	78.7%	75.6%	0.796
			H	12	32								
		VNA	L	129	43	0.727	0.750	0.766	0.404	0.862	74.5%	73.9%	0.796
			H	12	32								
		VNFB	L	129	53	0.761	0.709	0.745	0.387	0.842	71.9%	73.5%	0.798
			H	11	35								
		VNFBA	L	151	31	0.652	0.830	0.804	0.437	0.845	79.4%	74.1%	0.798
			H	16	30								
BPRS-Total	24	A	L	45	30	0.600	0.702	0.672	0.293	0.643	66.7%	65.1%	0.653
			H	42	99								
		VN	L	52	23	0.693	0.780	0.753	0.463	0.812	75.0%	73.7%	0.653
			H	31	110								
		VNA	L	54	21	0.720	0.773	0.758	0.480	0.809	75.5%	74.7%	0.653
			H	32	109								
		VNFB	L	56	24	0.700	0.723	0.720	0.408	0.776	71.5%	71.1%	0.649
			H	41	107								
		VNFBA	L	52	28	0.650	0.818	0.758	0.469	0.776	75.9%	73.4%	0.649
			H	27	121								
	32.0	A	L	82	56	0.594	0.500	0.567	0.091	0.561	56.0%	54.7%	0.639
			H	39	39								
		VN	L	88	50	0.638	0.833	0.714	0.453	0.772	70.8%	73.6%	0.639
			H	13	65								
		VNA	L	100	38	0.725	0.667	0.708	0.381	0.737	70.4%	69.6%	0.639
			H	26	52								
		VNFB	L	114	32	0.781	0.561	0.700	0.346	0.740	70.2%	67.1%	0.640
			H	36	46								
		VNFBA	L	84	62	0.575	0.707	0.630	0.272	0.716	62.3%	64.1%	0.640
			H	24	58								

TABLE 6.5: Results for automated prediction of NSA16 indices using verbal (V), non-verbal (N), facial (F), body movement (B), and speech adaptation (A) cues.

Scale	THR	Modality	CM			SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
				Predicted									
				L	H								
NSA2	2	A	L	87	30	0.744	0.636	0.693	0.382	0.756	69.4%	69.0%	0.542
			H	36	63								
		VN	L	87	30	0.744	0.788	0.764	0.530	0.818	76.4%	76.6%	0.542
			H	21	78								
		VNA	L	88	29	0.752	0.747	0.750	0.498	0.831	75.0%	75.0%	0.542
			H	25	74								
		VNFB	L	90	32	0.738	0.802	0.768	0.538	0.821	76.8%	77.0%	0.535
			H	21	85								
		VNFBA	L	95	27	0.779	0.764	0.772	0.542	0.841	77.2%	77.1%	0.535
			H	25	81								
NSA3	3	A	L	57	73	0.438	0.744	0.557	0.186	0.608	56.0%	59.1%	0.602
			H	22	64								
		VN	L	79	51	0.608	0.709	0.652	0.311	0.651	64.8%	65.8%	0.602
			H	25	61								
		VNA	L	79	51	0.608	0.698	0.647	0.299	0.660	64.4%	65.3%	0.602
			H	26	60								
		VNFB	L	80	55	0.593	0.688	0.635	0.276	0.635	63.2%	64.0%	0.592
			H	29	64								
		VNFBA	L	82	53	0.607	0.677	0.639	0.280	0.643	63.6%	64.2%	0.592
			H	30	63								
NSA6	3	A	L	91	36	0.717	0.584	0.662	0.301	0.669	66.2%	65.0%	0.588
			H	37	52								
		VN	L	82	45	0.646	0.831	0.724	0.472	0.722	72.2%	73.9%	0.588
			H	15	74								
		VNA	L	88	39	0.693	0.753	0.720	0.439	0.734	71.8%	72.3%	0.588
			H	22	67								
		VNFB	L	92	39	0.702	0.711	0.708	0.410	0.730	70.6%	70.7%	0.575
			H	28	69								
		VNFBA	L	89	42	0.679	0.794	0.729	0.468	0.735	72.8%	73.7%	0.575
			H	20	77								
NSA9	2	A	L	126	28	0.818	0.500	0.725	0.323	0.751	72.7%	65.9%	0.713
			H	31	31								
		VN	L	107	47	0.695	0.758	0.726	0.413	0.772	71.3%	72.6%	0.713
			H	15	47								
		VNA	L	133	21	0.864	0.694	0.816	0.552	0.803	81.5%	77.9%	0.713
			H	19	43								
		VNFB	L	108	53	0.671	0.716	0.698	0.355	0.760	68.4%	69.4%	0.706
			H	19	48								
		VNFBA	L	130	31	0.807	0.701	0.781	0.489	0.786	77.6%	75.4%	0.706
			H	20	47								
NSA-AB	5.301	A	L	74	35	0.679	0.720	0.699	0.399	0.724	69.9%	69.9%	0.505
			H	30	77								
		VN	L	85	24	0.780	0.729	0.754	0.510	0.802	75.5%	75.4%	0.505
			H	29	78								
		VNA	L	77	32	0.706	0.813	0.759	0.522	0.813	75.9%	76.0%	0.505
			H	20	87								
		VNFB	L	89	25	0.781	0.772	0.776	0.553	0.813	77.6%	77.6%	0.500
			H	26	88								
		VNFBA	L	82	32	0.719	0.842	0.780	0.566	0.820	78.1%	78.1%	0.500
			H	18	96								

TABLE 6.5: Results for automated prediction of NSA16 indices using verbal (V), non-verbal (N), facial (F), body movement (B), and speech adaptation (A) cues.

Scale	THR	Modality	CM			SEN	SPE	F1	MCC	AU-PRC	ACC	BAC	MB
				Predicted									
				L	H								
NSA-AM	8.454	A	L	71	36	0.664	0.440	0.545	0.107	0.546	55.1%	55.2%	0.505
			H	61	48								
		VN	L	79	28	0.738	0.587	0.660	0.329	0.704	66.2%	66.3%	0.505
			H	45	64								
		VNA	L	50	57	0.467	0.807	0.628	0.292	0.676	63.9%	63.7%	0.505
			H	21	88								
		VNFB	L	81	32	0.717	0.626	0.670	0.344	0.707	67.1%	67.1%	0.504
			H	43	72								
		VNFBA	L	80	33	0.708	0.600	0.653	0.310	0.685	65.4%	65.4%	0.504
			H	46	69								
NSA-PQ	2.549	A	L	74	39	0.655	0.583	0.620	0.238	0.642	62.0%	61.9%	0.523
			H	43	60								
		VN	L	92	21	0.814	0.583	0.699	0.409	0.723	70.4%	69.8%	0.523
			H	43	60								
		VNA	L	78	35	0.690	0.718	0.704	0.408	0.732	70.4%	70.4%	0.523
			H	29	74								
		VNFB	L	91	26	0.778	0.586	0.681	0.371	0.708	68.4%	68.2%	0.513
			H	46	65								
		VNFBA	L	80	37	0.684	0.721	0.702	0.404	0.722	70.2%	70.2%	0.513
			H	31	80								
NSA-RS	2.819	A	L	72	42	0.632	0.765	0.694	0.398	0.760	69.4%	69.8%	0.528
			H	24	78								
		VN	L	91	23	0.798	0.725	0.763	0.526	0.796	76.4%	76.2%	0.528
			H	28	74								
		VNA	L	86	28	0.754	0.755	0.755	0.509	0.812	75.5%	75.5%	0.528
			H	25	77								
		VNFB	L	85	33	0.720	0.809	0.763	0.530	0.793	76.3%	76.5%	0.518
			H	21	89								
		VNFBA	L	90	28	0.763	0.755	0.759	0.517	0.814	75.9%	75.9%	0.518
			H	27	83								

adaptation feature sets with VN feature sets and VNFB feature sets, the prediction results of the BACS-Total scale dropped 1.7% to 5.5%. Moreover, for predicting the severity of general psychiatric symptoms assessed by BPRS, including speech adaptation feature sets improves the prediction results between symptomatic and asymptomatic (BPRS-Total cut-off threshold is 24); however, it does not help distinguish between mild and severe symptoms (BPRS-Total cut-off threshold is 32).

To explore whether speech adaptation features could be used to predict individual and factor scales of NSA-16, the same ensemble learning method with LOOCV was applied in this chapter (refer to Section 3.6), and the results were presented in Table 6.5. From the results, we note that the prediction results of many NSA scores are slightly promoted after being incorporated into speech adaptation features.

This is especially the case for NSA9 (poor rapport), which can be distinguished with a BAC of 77.9% when using VNA feature sets. Intuitively, this fits the hypothesis that the speech adaptive measurements between the participant and psychiatrist can capture the interactive information, and this information can be used to predict the subjective clinical assessment rating evaluated by the psychiatrist.

6.4 Salient Features

In automated classification tasks, machine learning models mine and learn the most significant features and make predictions based on those important features. By analyzing important features, we can gain insight into what distinguishes the different participants and then interpret our findings. Therefore, I explored the most salient speech adaptation features to understand the differences in the vocal interaction produced by schizophrenic patients and healthy controls with psychiatrists. The Kruskal-Wallis test was applied to all speech adaptation features and computed the corresponding p-values with FDR post-hoc correction (see Section 3.8). The top 20 salient categories and their p-values are shown in Table 6.6.

When compare the speech adaptation features of patients with schizophrenia to healthy controls, we note that the loudness (50th percentile) of patients with schizophrenia has a significantly higher correlation with the standard deviation of the center frequency of first (0.0026, 95% CI - -0.009 to 0.014; $P<0.05$) and second formant (-0.0267, 95% CI - -0.040 to -0.014; $P<0.05$) of psychiatrists. Moreover, the correlation between the average F0 falling slope of schizophrenia patients and the third coefficient of MFCC of psychiatrists is higher than that of healthy controls. Although this correlation feature does not exhibit differences that remain statistically significant after post-correction, these features ranked high in distinguishing schizophrenia from healthy controls. Apart from the correlation-based features, patients with schizophrenia have significantly higher similarity (1 - SMAPE) between the STD of third formant frequency and the average jitter of the psychiatrist (0.2575, 95% CI - 0.2523 to 0.2627; $P<0.05$). In addition, the similarity between the first formant bandwidth STD of schizophrenia patients and the second formant bandwidth STD of psychiatrists is also different from that of healthy participants (0.2592, 95% CI - 0.2553 to 0.2630; $P<0.05$). These observations have not been

TABLE 6.6: Top 10 salient speech adaptation features in pairwise classification tasks.

Task	Method	Participant's feature<Method>Psychiatrist's feature
$\overline{S}$ vs. H	SMAPE	**F0_FSlope_std<SMAPE>UVSL_std**
		HD1-2_mean<SMAPE>Jitter_mean
		FormantFreq3_std<SMAPE>LoudnessPeak
		F0_RSlope_std<SMAPE>UVSL_std
		FormantBand3_std<SMAPE>VS
		MFCC3_mean<SMAPE>Jitter_mean
		FormantFreq3_std<SMAPE>Jitter_mean*
		FormantBand1_std<SMAPE>FormantBand2_std*
		FormantBand2_mean<SMAPE>SpecFlux_mean
		MFCC4_mean<SMAPE>SpectralSlopeH_mean
	Correlation	**F0_FSlope_mean<reciprocity>MFCC3_mean**
		Loudness_pct50<reciprocity>FormantFreq1_std*
		Loudness_pct0-20<reciprocity>FormantBand1_std*
		Loudness_pct50<reciprocity>FormantFreq2_std*
		MFCC4_mean<reciprocity>SpectralSlopeH_mean
		MFCC1_std<reciprocity>FormantBand1_std*
		MFCC3_std<reciprocity>MFCC4_std
		VS<reciprocity>VSL_std
		HD1-3_std<reciprocity>LoudnessPeak
		UVSL_std<reciprocity>RFN1_mean
	Convergence	**SpecFlux_std<convergence>MFCC3_mean**
		Shimmer_mean<convergence>HD1-2_mean
		Loudness_FSlope_std<convergence>Loudness_pct50
		HamIndex_std<convergence>MFCC4_mean
		FormantBand3_std<convergence>MFCC2_mean
		F0_FSlope_std<convergence>MFCC2_mean*
		FormantBand3_std<convergence>MFCC1_mean
		slopeV0-500_std<convergence>MFCC3_std
		F0_FSlope_std<convergence>FormantBand3_mean
		Shimmer_std<convergence>MFCC3_std
$\overline{D}$ vs. H	SMAPE	UVSL_std<SMAPE>Loudness_std
		VS<SMAPE>SpecFlux_std
		F0_std<SMAPE>Loudness_pct50
		SpecFlux_std<SMAPE>HamIndex_mean
		FormantFreq3_std<SMAPE>Loudness_RSlope_mean
		equivalentSoundLevel_dBp<SMAPE>LoudnessPeak
		F0_FSlope_mean<SMAPE>slopeV0-500_std
		HNR_std<SMAPE>F0_FSlope_std
		SpectralSlopeH_mean<SMAPE>SpectralSlopeH_mean
		Loudness_std<SMAPE>FormantBand1_std
	Correlation	**AlphaR_UV_mean<reciprocity>AlphaR_mean**
		SpectralSlopeH_mean<reciprocity>SpectralSlopeH_mean
		MFCC4_mean<reciprocity>MFCC4_mean
		F0_FSlope_mean<reciprocity>Loudness_pct50
		SpecFlux_std<reciprocity>AlphaR_mean
		UVSL_std<reciprocity>Loudness_FSlope_std
		UVSL_std<reciprocity>Loudness_RSlope_std
		HamIndex_mean<reciprocity>AlphaR_UV_mean
		MFCC3_mean<reciprocity>HD1-3_std
		MFCC4_std<reciprocity>UVSL_mean

TABLE 6.6: Top 10 salient speech adaptation features in pairwise classification tasks.

Task	Method	Participant's feature<Method>Psychiatrist's feature
$\overline{D}$ vs. H	Convergence	**Loudness_pct0-20<convergence>HNR_mean**
		UVSL_std<convergence>MFCC3_mean
		Loudness_RSlope_std<convergence>SpecFlux_mean
		FormantFreq2_std<convergence>FormantFreq1_std
		Jitter_mean<convergence>Loudness_pct50
		FormantBand1_mean<convergence>F0_mean
		FormantBand1_mean<convergence>F0_pct50
		HNR_std<convergence>MFCC2_mean
		UVSL_mean<convergence>Loudness_FSlope_std
		FormantFreq3_std<convergence>FormantBand2_std
$\overline{D}$ vs. S	SMAPE	FormantFreq1_std<SMAPE>UVSL_mean**
		Shimmer_mean<SMAPE>FormantBand3_mean*
		Loudness_std<SMAPE>VSL_std*
		RFN1_std<SMAPE>FormantFreq3_mean*
		VS<SMAPE>SpecFlux_mean
		VSL_mean<SMAPE>SpecFlux_mean*
		SpectralSlopeH_mean<SMAPE>SpecFlux_std
		AlphaR_std<SMAPE>FormantBand2_std**
		SpectralSlopeH_mean<SMAPE>HD1-2_mean*
		Loudness_pct20<SMAPE>RFN1_mean
	Correlation	SpecFlux_std<reciprocity>AlphaR_UV_mean
		MFCC1_std<reciprocity>SpectralSlopeH_mean*
		MFCC3_mean<reciprocity>HD1-3_std*
		Loudness_RSlope_mean<reciprocity>FormantFreq2_std**
		F0_RSlope_mean<reciprocity>F0_mean
		MFCC2_mean<reciprocity>Loudness_pct20*
		MFCC2_mean<reciprocity>FormantBand2_std*
		AlphaR_std<reciprocity>FormantBand2_std*
		SpectralSlopeH_mean<reciprocity>MFCC4_std
		F0_RSlope_mean<reciprocity>F0_pct80
	Convergence	**F0_FSlope_mean<convergence>UVSL_std***
		Loudness_FSlope_std<convergence>HD1-3_mean*
		F0_RSlope_std<convergence>MFCC4_mean*
		F0_pct0-20<convergence>F0_FSlope_mean*
		Shimmer_mean<convergence>F0_std
		RFN1_mean<convergence>Loudness_FSlope_std*
		VSL_std<convergence>Loudness_RSlope_mean
		F0_pct20<convergence>F0_mean
		FormantFreq1_std<convergence>Loudness_pct0-20*
		F0_pct50<convergence>SpectralSlopeH_mean*

Suffix: min-minimum; max-maximum; sk-skewness; ku-kurtosis; std-standard deviation.
**: p-value $\leq$ 0.005; *: p-value $\leq$ 0.05; Kruskal-Wallis tests with FDR correction.
bold feature: the average value of this feature is larger in the class with overline.

reported in the literature so far, consequently, speech adaptation cues deserve more attention in the context of understanding schizophrenia patients.

Interestingly, our system also captured various behavioral cues between depression and schizophrenia. For instance, the STD of the first formant frequency of patients

with schizophrenia is closer to the mean unvoiced segment length of psychiatrists than patients with depression (0.2591, 95% CI - 0.2543 to 0.264; P<0.005). In addition, the similarity between the variance of the alpha ratio for depression patients and the variance of the second formant bandwidth for psychiatrists is significantly higher than that for schizophrenia (0.2745, 95% CI - 0.2669 to 0.2821; P<0.005). For correlation-based features, the Pearson correlation between the average loudness rising slope of depression and the STD of the second formant central frequency of psychiatrists is significantly lower than schizophrenia (-0.0538, 95% CI - -0.0694 to -0.0381; P<0.005). Besides, compared with patients with schizophrenia, the STD of the average MFCC2 and alpha ratio of depression patients is more positively correlated with the STD of the second formant bandwidth of the psychiatrist. In addition, there are significant statistical differences in the convergence of various voice characteristics between schizophrenia and depression in the interview with psychiatrists. For instance, there is a higher level of convergence between the mean pitch falling slope and the STD of unvoiced segment length between depression patients and psychiatrists (0.4771, 95% CI - 0.4731 to 0.4812; P<0.05).

6.5 Conclusion

In conclusion, patients with schizophrenia present different characteristics of speech adaptation compared with healthy controls. These differences promote the machine learning model on objectively classifying patients with schizophrenia from healthy controls by about 3.4%. Besides, by using the speech adaptation features extracted from the participant-psychiatrist dialogue, the classification performance between depression and schizophrenia patients improves 1.5%. However, there were no statistically significant speech adaptation features between depression and healthy control groups after post-correction, which may be a reason why utilizing these characteristics can not promote the classification results of depression vs. healthy control (see Table 6.2). This might be the first study to investigate how well a machine learning pipeline along with bivariate speech adaptation features for schizophrenia and depression detection. For better understanding the core results of this chapter, we summarized the classification results and prediction results in Table 6.7.

TABLE 6.7: Summarized results for automated classification of schizophrenia, depression, and healthy controls using audiovisual and adaptive modalities.

Mode	Modality	Classification tasks				Prediction tasks					
						Negative symptoms		Cognitive symptoms		General psychiatric symptoms	
		S vs. H	D vs. H	D vs. S	DS vs. H	Global score	Total score	Normal vs. Mild	Mild vs. Severe	Normal vs. Mild	Mild vs. Severe
Single modality	Correlation	0.705	0.660	0.676	0.664	0.669	0.661	0.605	0.592	0.659	0.520
	SMAPE	0.642	0.659	0.727	0.655	0.620	0.649	0.649	0.612	0.619	0.575
	Convergence	0.617	0.525	0.640	0.532	0.518	0.537	0.521	0.606	0.485	0.526
Fusion of modalities	Adaptation	0.702	0.654	0.732	0.683	0.686	0.634	0.617	0.626	0.640	0.566
	Speech + Adaptation	0.839	**0.807**	0.819	**0.777**	0.732	0.752	**0.748**	**0.738**	**0.754**	**0.698**
	Speech + Facial + Movement + Adaptation	**0.841**	0.797	**0.848**	0.767	**0.746**	**0.756**	0.745	0.733	0.727	0.662

Adaptation modality: Convergence, SMAPE, and Correlation.
Facial modality: Affectiva, OpenFace, and Opsis.
Speech modality: LIWC, Diction, LDA, Doc2Vec, Conversational, DisVoice, OpenSmile, Affectiva, OpenFace, and Opsis.

To the best of our knowledge, there is also currently no research linking speech adaptation cues to the severity of symptoms in patients with mental illness. This thesis demonstrates that combining the speech adaptation features with audio-based features could predict NSA9 (poor rapport) with a BAC of 77.9%. NSA9 assesses the interviewer's subjective sense that he/she and the subject are actively engaged in communication with one another. Moreover, various salient speech adaptation features are observed between different groups of participants (schizophrenia, depression, and healthy controls). These characteristics are statistically significantly different after FDR post-correction. Therefore, these conversational cues deserve more attention in the context of negative symptoms and interaction assessments. In the future, these automated measurements can easily be implemented in a lightweight mobile app for monitoring patients over a long period from their phone call recordings.

Chapter 7

Discussion

Inspired by earlier promising studies of digital phenotyping of psychiatric patients, this thesis explored the relevance of a comprehensive portfolio of behavioral cues and signals extracted with state-of-the-art tools from the fields of signal processing and artificial intelligence, for differential diagnosis and for detecting psychiatric symptoms. This thesis also demonstrated that machine learning models could leverage various behavioral cues for differential diagnosis and reliable detection of various psychiatric symptoms. Furthermore, the proposed ensemble learning pipeline modularizes the different types of behavioral cues. Thus, in the future, additional behavioral cues or different kinds of digital biomarkers (e.g., number of daily steps) could be readily integrated into the system. In the following, our specific contributions in light of the quantitative results presented in the previous section and results available in the literature are discussed.

For classifying schizophrenia patients vs. healthy controls, the methods proposed in the literature achieved an accuracy between 70% and up to 90% (Table B.1 in the appendix B). However, since those studies considered different data sets, a direct comparison is not possible. Furthermore, the studies that achieved a high accuracy (close to 90%) are often limited to a small number of patients [35, 230], did not perform cross-validation [217], or strongly optimized the classifier at the risk of overfitting [39]. Therefore, those results might not be reliable. Furthermore, except for our previous preliminary studies [117], all existing studies for schizophrenia assess a single type of behavioral cue (e.g., prosodic cues). In contrast, our experiment recruited a larger number of schizophrenic patients (N=103) compared with

the existing literature. In an attempt to create a robust classification pipeline, five common classifiers were combined instead of relying on only a single classifier, did not optimize the parameters of the classifiers but chose the standard settings instead, and integrated multiple types of behavioral cues. The proposed pipeline generated numerous "votes" from each component classifier and for each kind of behavioral signal. Next, the system made a decision (e.g., "schizophrenia" or "HC") based on majority voting. Such distributed design of classification pipelines, known as ensemble classification, naturally leads to more robust decisions.

For depression detection, the release of multiple multi-sensor open-source data sets of depression assessments (e.g., DAIC, BackDog, and Pitt) has led to significant research interest [50]. Most studies report ACC values between 70% and 95% (Table B.1 in the appendix B). However, it is again not straightforward to compare our study with those studies since the patients and the severity of the symptoms are different. Also, the control group in the DAIC data set exhibited PTSD symptoms, whereas the control group in the current study comprises healthy subjects [163, 189]. In the studies that report high accuracy values, the choice of behavioral cues and classifiers typically appear intensely optimized [54, 163, 225]. Although the proposed classification pipelines are not optimized in an attempt to achieve 90% accuracy, to avoid overfitting, it still showcases that the accuracy of the classification of depression and healthy controls in this study (BAC=82.3%) generally confirms the effectiveness of digital behavioral cues for distinguishing depression patients from HC.

Besides identifying patients from healthy controls, our findings also support the transdiagnostic view of clinical symptom manifestations outlined by RDoC [59]. The proposed pipeline is able to differentiate depression and schizophrenia at an outstanding performance of ACC=85.6% (Table 5.1). This result compares favorably with the small number of earlier related studies: Lott et al. extracted linguistic features from the manual transcriptions of structured clinical interviews and designed models that, with those features as input, can differentiate among schizophrenia, depression, and bipolar disorders at an ACC of 72.7% [57], while Kliper et al. proposed models that are able to automatically differentiate the speech from schizophrenia and depression patients with an ACC range of 52.0% to 76.7% using acoustic features [58]. Several subtle quantitative differences in patients' behaviors that distinguish their clinical diagnoses (depression vs. schizophrenia),

such as the difference of turn-takings, use of verbs, eye closure variability, and Mel-frequency cepstral coefficient (MFCC); more information is referred to Table 4.4 and Table 6.6.

This thesis also explored the possibility of detecting a series of symptoms, namely negative, cognitive, and general psychiatric symptoms. Similar studies that report classification and regression results of negative, cognitive, and general psychiatric symptoms using machine learning techniques are summarized in Table B. For negative symptoms, we noted that our pipeline can predict negative symptoms related to diminished expression (e.g., NSA-RS, BAC=77.8%; NSA-AB, BAC=78.0%, Table 5.4) better when compared to diminished motivation (e.g., NSA-AM, BAC= 67.7%). Furthermore, the prediction result of the diminished expression (DE) and social amotivation (SA) domain score using PANSS (PANSS-DE, BAC=83.5%; PANSS-AM, BAC=65.7%; Table 5.7) for schizophrenia further supports this observation. The underlying reason for this might be that the behavioral cues collected in this study capture people's behavior and expression but have little to do with people's thoughts and motivations. In addition, we noted that the speech-related feature sets (verbal and nonverbal speech feature sets) are the most informative for detecting negative symptoms (Table 5.4), which is consistent with earlier observations that vocal expressions are statistically significantly correlated with negative symptom measures, especially restricted speech and affective blunting [55, 259]. A similar conclusion was reported by Cohen et al., who observed a strong correlation between speech cues and the severity of negative symptoms assessed by the Brief Negative Symptom Scale (BNSS) [27]. Next, Cohen designed a model that, with 138 acoustic features as input, can predict blunted affect and alogia scores measured by the Scale for the Assessment of Negative Symptoms (SANS) [55]. Since their data is unbalanced, the BAC based on the metrics they provided was calculated for a fair comparison. Their results (blunted affect: BAC=78.5%; and alogia: 81.0%) are in line with ours (NSA-AB, BAC=78.0%; NSA-RS, BAC=78.8%, Table 5.4).

To the best of our knowledge, this thesis is the first to propose automated methods for predicting the severity of cognitive symptoms for schizophrenia or depression. The proposed system can detect mild to severe (BAC=78.9%, Table 5.5) cognitive symptoms as well as severe (BAC=75.6%, Table 5.5) cognitive symptoms for all three groups of subjects combined, with similar prediction results for each patient

group separately. Moreover, we noted that the proposed pipeline could also accurately predict BACS-TMT and BACS-SC (Table 5.5). The BACS-TMT and BACS-SC scores assess the patient's processing speed and problem-solving ability and are highly correlated with the expression domain for schizophrenia [13]. Again, these results indicate that audio-visual behavioral characteristics are able to predict clinical ratings related to expression levels. However, more research is warranted to explore the common neural substrates between expression mechanisms and cognition for schizophrenia and depression. In the long term, automated detection of cognitive symptoms may overcome some of the shortcomings of conventional assessments. For instance, BACS requires half an hour for a single standard battery of tests [237], which could be avoided by automated prediction of BACS from short audio-visual recordings (e.g., phone calls).

For general psychiatric symptoms, the proposed model was showcased that it is able to classify BPRS-Total rating on the three groups combined, with solid results on the negative symptom factor score of BPRS but relatively poor results on the positive, affective, and resistance factor scores (Table 5.6). Moreover, the proposed system could predict PANSS more reliably than BPRS (see Table 5.6 and Table 5.7). The underlying reason for this difference might be that the evaluation of symptoms by PANSS is more extensive than BPRS, as PANSS contains more expression-related scales, such as the lack of spontaneity and flow of conversation [119]. There are a few studies in the literature; however, these studies are not directly comparable as the objectives are different: Tron et al. extracted facial features of 34 patients with schizophrenia and 20 healthy controls, and showcased that several PANSS ratings were moderately correlated with the regression outputs [37]. Wörtwein et al. relied on acoustic features to predict the total BPRS scores of 20 psychiatric patients at an average absolute error of 10.35 [271]. The results from these two studies combined with the present study suggest that predicting general psychiatric symptoms from audio-visual behavioral cues is a promising avenue for future research.

In summary, when testing on the data of all groups combined, the BAC of the proposed machine learning system reached a moderate-high level (75% to 85%) on all total symptom severity scores (e.g., NSA-Total, BACS-composite, BPRS-Total, and PANSS-Total) and some factors/subscales (e.g., NSA-RS, NSA-AB, BACS-TMT, BPRS-NEG, PANSS-NSFS, and PANSS-DE); a moderate level (65% to

75%) on NSA-AM, BACS-SC, BACS-VM, BPRS-AFF, and BPRS-POS; and a poor level (<65%) on BACS-SF and BPRS-RES. In addition, as previously discussed, the proposed pipeline achieves a moderate-high prediction accuracy for the clinical scales related to diminished expression (except for BACS-SC, for which achieved only moderate results), while it performs rather poorly on other scales.

Chapter 8

Conclusion

8.1 Conclusion

This pilot study collected the interview recordings from 50 patients with depression, 103 patients with schizophrenia, and 75 healthy controls. We build an ensemble learning pipeline to automatically classify different groups of participants and predict the severity of different patients' symptoms (negative, cognitive, and general psychiatric symptoms). Our methods integrate the analysis results of language usage, non-verbal cues, and adaptation in speech, facial expression, and body movement and provide salient interpretable features to clarify the clues and trends learned by machine learning algorithms. In particular, these results are promising and are an essential step towards our overall goal of creating an automated system to help clinical diagnosis, evaluation, and monitoring of schizophrenia and depression.

Many exciting results and interesting trends were observed in my analysis. For example, the proposed machine learning system achieves a moderate-high accuracy for classifying the total score of negative symptoms (BAC=76.0%; SEN=80.2%; SPE=71.8%), the composite score of cognitive symptoms (BAC=75.6%; SEN=80.8%; SPE=70.5%), and total score of general psychiatric symptoms (BAC=73.6%; SEN=83.3%; SPE=63.8%). Our results particularly demonstrate the success of predicting assessment ratings that are directly or indirectly related to diminished expressions with a moderate-high balanced accuracy (>75%), such as restricted speech, affective blunting, and token motor test, while achieving relatively poor

results (<65%) on semantic fluency and resistance factor scores, which are not directly related to diminished expression. The speech adaptation cues are good indicators to differentiate the clinical rating on poor rapport assessed by NSA-16. Furthermore, the proposed system can differentiate schizophrenia and depression recordings from healthy control recordings with 85.2 and 82.3% balanced accuracy, respectively, differentiate between depression and schizophrenia with a balanced accuracy of 84.7%, and distinguish the three groups combined (schizophrenia, depression, and healthy controls) with a 3-class classification accuracy of 68.7%.

There is a strong unmet need for technologies capable of monitoring mental health conditions over a long period of time. Due to their heavy workload, clinicians cannot frequently contact their patients, and the recent covid-19 pandemic has only exacerbated this problem [272, 273]. The findings in the present study strongly support the potential of developing digital phenotyping pipelines that leverage a comprehensive set of audio-visual behavioral traits, potentially combined with other kinds of measurements (e.g., mobility patterns, vital signals), for accurate and unbiased long-term psychiatric assessment that can be used remotely. Such a pipeline may provide valuable early diagnosis and longitudinal monitoring of severe mental illness if implemented as a smartphone app or a virtual healthcare application. In the future, we will be expanding the data sets to cover participants from multiple institutions and with a wide variety of cultural and ethnic backgrounds. We also plan to collect mobile calls with proper consent and develop similar machine learning pipelines on unstructured data. Ultimately, we hope that this research direction will contribute to developing a low-cost platform that provides unbiased psychiatric assessment and monitoring for people with severe mental illness while respecting the individual's privacy and the proper life cycle management of personal data.

8.2 Limitation

This study has the following limitations. First, our proposed machine learning pipeline is modular, and we achieved high performance in predicting the scores related to diminished expression. However, certain assessment scores, such as anhedonia, avolition, and asociality domains, were harder to predict through our digital features. This problem may be because we deal with patients presenting

mild symptoms, and some symptoms were absent. Therefore, it is important to utilize other measurements, such as Ecological Momentary Assessment (EMA), GPS tracker, and vital signals, to complement our current pipeline and provide a more comprehensive view of psychiatric symptoms.

Since most of the patients involved in this study exhibited mild symptoms, the machine learning algorithms could not thoroughly learn the characteristics of more severe cases when predicting the disease's severity. Therefore, it is vital to develop multi-center datasets to enlarge the sample size and balance symptom severity distribution. Moreover, all participants in both studies were of Asian ethnicity. Therefore, it is necessary to validate our models in populations with diverse ethnicities and cultures in future studies. Furthermore, the automatic speech recognition and facial analysis tools used in this study were trained on data collected in the United States; hence, they may perform less reliably on the present study's data [274].

Finally, the data for the present study were collected during three visits over a period of only 12 weeks. To realize the full potential of long-term digital phenotyping on smartphones for longitudinal follow-up of psychiatric patients, long-term data collection of a larger group of patients will be required over a longer period of time.

8.3 Future Work

This thesis has explored the feasibility of using objective audio-visual features to assess mental illness patients' behavior. In this chapter, some of the future work are outlined.

8.3.1 Semi-structured to Unstructured Interviews

Highly trained psychiatrists conduct current experiments outlined in this report in an indoor environment. The interviews with mentally ill patients follow the same protocol, with a fixed set of guidelines. A list of discussion prompts is associated with each topic area and contains both open-ended and closed-ended questions. However, a lightly structured or unstructured interview communicates with interviewers in a more neutral environment with less attached bias compared with the

semi-structured interview. Therefore, the characteristics of the unstructured interview may vary from one conversation to another. Nevertheless, it will further give us ideas on how our algorithms work in the daily interactions of patients.

8.3.2 Explore the Measurement of Social Amotivation

Compared with normal people, people with schizophrenia have reduced social drive, sense of purpose, and daily activity [17]. As mentioned before, our pipeline can predict symptoms related to diminished expression better when compared to social amotivation. The underlying reason for this might be that the behavioral cues collected in this study capture people's behavior and expression but have little to do with people's thoughts and motivations. Therefore, it is important to utilize other measurements, such as Ecological Momentary Assessment (EMA), GPS tracker, and vital signals.

EMA is a method that measures the subject's behavior, experience, and response in their natural environment. It can be embedded in a smartphone and has been used to measure the real-world functioning, motivational negative symptoms, and stress reactivity of patients with schizophrenia [275]. Moreover, GPS mobility is also a promising digital biomarker for measuring the behaviors of mental disorders. It has been found that less GPS mobility is associated with more serious negative symptoms, especially reduced motivation and purpose, while greater GPS mobility is associated with more social functioning [276]. Therefore, integrating the analysis using EMA and GPS may complement the shortcoming of the current pipeline and be able to evaluate the negative symptoms and the general psychiatric symptoms of mental disorders more comprehensively.

8.3.3 Explore the Phenomena of Interaction Coordination

Adaptation is a biological and social behavior for human beings, and it plays a fundamental role in our daily activities. Many researchers have summarized this phenomenon from different angles (e.g., CAT [96] and IAT [101]). Apart from the univariate analysis (e.g., verbal and non-verbal features), this thesis quantified the conversational features and similarity, reciprocity, and convergence between the acoustic and prosodic features of the two interlocutors. However, since only audio

recordings of the psychiatrist and the patient were collected in the experiment, the audio-visual interaction needed to be analyzed in the future.

Apart from speech-related interaction, audio-visual interaction is also an interesting research direction that may need attention. In social interaction, the conversation partner's auditory and visual sensory information and the environment can be spontaneously integrated to construct inherent behavioral characteristics. Researchers have found that autism spectrum disorder and schizophrenia could lead to abnormal sensory experiences and social communication disorders [277]. In particular, studies have found that the temporal binding window (TBW) of patients with mental illness is different when facing different audio-visual signals. Mental disorders with schizophrenia or autism often feel that some audiovisual signals separated by a long time still come from the same event [278], which may lead to abnormal cognition, perception, and language expression [279, 280]. Therefore, by automatically quantifying the relevance of speech and facial expressions between dialogues, it is feasible to capture different responses to different auditory and visual signals, thereby facilitating the diagnosis and assessment of people with mental illness.

8.3.4 Long-term Monitoring App Design

As mentioned earlier, Study-A collected recordings of all participants only three times over a 12-week period. To fulfill the goal of long-term monitoring of mental disorders, the next step for data collection would be to develop software, including an app-based system, for eventual clinical validation and subsequent implementation. Upon completing this study, we envisage using the developed algorithms and pipelines in this thesis to translate speech and visual characteristics into clinically meaningful information on negative symptoms, cognitive and social-cognitive profiles. With the widespread adoption of this system, we believe a larger proportion of under-served people who suffer from these symptoms and deficits might be able to access a suite of psychosocial services that might aid in their psychosocial rehabilitation and functional recovery. This is an important first step for us to develop local treatment protocols and regimens to remediate these impairments.

Appendix A

Prediction results of NSA-16 indices.

TABLE A.1: Results for automated prediction of NSA-16 scales for schizophrenia (S), depression (D), and healthy controls (H) using audio-visual cues.

Sample	Score	THR	Feature	CM			SEN	SPE	F1	AU-PRC	ACC	BAC	MB
					Predicted								
					L	H							
S	NSA1	2.00	F	L	25	9	0.700	0.735	0.746	0.803	0.727	0.718	0.773
				H	3	7							
	NSA2	2.00	N	L	35	10	0.792	0.778	0.786	0.842	0.786	0.785	0.541
				H	11	42							
	NSA3	2.00	VN	L	40	23	0.829	0.635	0.710	0.764	0.704	0.732	0.643
				H	6	29							
	NSA6	3.00	VNF	L	33	12	0.741	0.733	0.738	0.703	0.737	0.737	0.545
				H	14	40							
	NSA10	4.00	B	L	34	12	0.674	0.739	0.706	0.725	0.707	0.707	0.500
				H	15	31							
	NSA15	3.00	VNF	L	40	14	0.689	0.741	0.717	0.696	0.717	0.715	0.545
				H	14	31							
D	NSA1	2.00	B	L	27	3	0.583	0.900	0.804	0.799	0.810	0.742	0.714
				H	5	7							
	NSA7	4.00	V	L	23	5	0.700	0.821	0.770	0.794	0.771	0.761	0.583
				H	6	14							
	NSA8	4.00	N	L	13	3	0.750	0.813	0.777	0.827	0.771	0.781	0.667
				H	8	24							
	NSA9	2.00	N	L	24	4	0.750	0.857	0.812	0.713	0.813	0.804	0.583
				H	5	15							
	NSA10	3.00	F	L	18	4	0.650	0.818	0.736	0.744	0.738	0.734	0.524
				H	7	13							
	NSA15	3.00	VNF	L	24	3	0.609	0.889	0.754	0.736	0.760	0.749	0.540
				H	9	14							

TABLE A.1: Results for automated prediction of NSA-16 scales for schizophrenia (S), depression (D), and healthy controls (H) using audio-visual cues.

Sample	Score	THR	Feature	CM			SEN	SPE	F1	AU-PRC	ACC	BAC	MB
					Predicted								
					L	H							
DS	NSA1	2.00	F	L	48	16	0.909	0.750	0.803	0.818	0.791	0.830	0.744
				H	2	20							
	NSA2	2.00	VNF	L	54	16	0.759	0.771	0.765	0.810	0.765	0.765	0.530
				H	19	60							
	NSA6	3.00	F	L	31	7	0.646	0.816	0.721	0.688	0.721	0.731	0.558
				H	17	31							
	NSA15	3.00	N	L	59	21	0.667	0.738	0.705	0.726	0.705	0.702	0.548
				H	22	44							
SH	NSA1	2.00	VNFB	L	96	42	0.750	0.696	0.730	0.785	0.708	0.723	0.775
				H	10	30							
	NSA2	2.00	VN	L	58	35	0.947	0.624	0.765	0.850	0.768	0.785	0.554
				H	4	71							
	NSA3	2.00	VNFB	L	64	11	0.573	0.853	0.689	0.705	0.691	0.713	0.579
				H	44	59							
	NSA5	2.00	VN	L	96	27	0.822	0.780	0.801	0.845	0.792	0.801	0.732
				H	8	37							
	NSA6	2.00	F	L	24	10	0.800	0.706	0.765	0.790	0.764	0.753	0.618
				H	11	44							
	NSA14	3.00	VN	L	50	16	0.706	0.758	0.729	0.729	0.726	0.732	0.607
				H	30	72							
	NSA15	2.00	VNFB	L	64	10	0.721	0.865	0.782	0.790	0.781	0.793	0.584
				H	29	75							
DH	NSA2	2.00	VNFB	L	55	18	0.788	0.753	0.769	0.760	0.768	0.771	0.584
				H	11	41							
	NSA5	3.00	N	L	13	4	0.760	0.765	0.764	0.771	0.762	0.762	0.595
				H	6	19							
	NSA6	2.00	N	L	37	20	0.902	0.649	0.776	0.827	0.780	0.775	0.517
				H	6	55							
	NSA9	2.00	VN	L	68	18	0.656	0.791	0.761	0.775	0.754	0.723	0.729
				H	11	21							
	NSA14	3.00	VNFB	L	51	14	0.683	0.785	0.735	0.708	0.736	0.734	0.520
				H	19	41							
	NSA15	2.00	VNF	L	42	15	0.746	0.737	0.742	0.753	0.742	0.742	0.540
				H	17	50							
	NSA16	2.00	B	L	41	27	0.875	0.603	0.707	0.744	0.704	0.739	0.630
				H	5	35							
DSH	NSA2	2.00	VNFB	L	90	32	0.802	0.738	0.768	0.821	0.768	0.770	0.535
				H	21	85							
	NSA6	3.00	VN	L	82	45	0.831	0.646	0.724	0.722	0.722	0.739	0.588
				H	15	74							
	NSA8	4.00	VNFB	L	109	23	0.615	0.826	0.733	0.727	0.737	0.720	0.579
				H	37	59							
	NSA9	2.00	VN	L	107	47	0.758	0.695	0.726	0.772	0.713	0.726	0.713
				H	15	47							
	NSA14	4.00	VNFB	L	94	35	0.697	0.729	0.716	0.742	0.715	0.713	0.566
				H	30	69							
	NSA15	2.00	VNFB	L	68	19	0.752	0.782	0.766	0.798	0.763	0.767	0.618
				H	35	106							

Appendix B

Related Work.

TABLE B.1: Overview of related behavioral data-driven studies on diagnosis of schizophrenia and depression.

Task	Aspect	Reference	Year	Dataset	Data Source	Features	Classifier	Performance (ACC/BAC/AUC)
SCZ vs. HC	Verbal (spoken language)	Hong et al,[115]	2015	SCZ = 23 (12M, 11F); HC = 16 (7M, 9F)	Manual transcriptions of emotional narrative	Lexical features	SVM	ACC = 74.4% (with CV)
		Elvevåg B, et al.[214]	2007	SCZ = 26 (19M, 7F); HC = 25 (10M, 15F)	Manual transcriptions of clinical interview	LSA-based features	LDA	ACC=78.4% (with CV)
		Elvevåg et al.[215]	2010	SCZ = 53 (19M, 7F); HC = 30 (10M, 15F)	Manual transcriptions of clinical interview	LSA-based features	LDA	ACC=77.1% (with CV)
		Corcoran et al.[126]	2018	SCZ (CHR+) = 19 (17M, 2F); HC = 21 (13M, 8F)	Manual transcriptions of Caplan's "Story Game"	LSA-based features	LR	ACC=72% (with CV)
		Bar et al.[127]	2019	SCZ = 24; HC = 27 (All Male)	Manual transcriptions of 18 clinical questions	Semantic features	RF, XGB, and SVM	ACC=70.4 to 81.5% (with CV)
		Mota et al.[230]	2012	SCZ = 8 (7M, 1F); HC = 8 (5M, 2F); Manic = 8 (8M, 0F)	Manual transcriptions of the description of a recent dream	Graph-based features	NB, SVM, DT, MLP, and RBF	AUC=50 to 90% (with CV)
		Rezaii et al.[217]	2019	Training set: SCZ = 7 (4M, 3F); HC = 23 (10M, 15F) Testing set: SCZ = 5 (19M, 7F); HC = 5 (10M, 15F)	Manual transcriptions of structured interview	Word2Vec-based semantic density	NA	Training: ACC=93%; Testing: ACC=90%
		Xu et al.[30]	2018	SCZ = 50 (25M, 25F); HC = 25 (11M, 14F)	Kaldi automated transcriptions of semi-structured clinical interview	LIWC, DICTION7.0, Doc2Vec	Ensemble classifier	ACC=78.7% (with CV)
		Xu et al.[281]	2019	SCZ = 47 (22M, 25F); HC = 24 (10M, 14F)	Google ASR automated transcriptions of semi-structured clinical interview	LIWC	SVM, LR, KNN	ACC=84.5 to 85.9% (with CV)
		Tang et al.[128]	2021	SCZ = 20 (11M, 9F); HC = 11 (4M, 7F)	Manual transcriptions of semi-structured clinical interview	BERT-derived features	NB	ACC = 87%, AUC=0.91 (with CV)

TABLE B.1: Overview of related behavioral data-driven studies on diagnosis of schizophrenia and depression.

Task	Aspect	Reference	Year	Dataset	Data Source	Features	Classifier	Performance (ACC/BAC/AUC)
SCZ vs. HC	Non-verbal (acoustic/ prosodic/ conversational)	Kliper et al.[58]	2019	SCZ = 22; DP = 20; HC = 20	NA	Acoustic features	Linear Classifier	SCZ vs. HC: ACC=73.8 to 81.0% DP vs. HC: ACC=69.0 to 87.1% SCZ vs. DP: ACC=52.0 to 76.7% (all with CV)
		Espinola et al.[58]	2021	SCZ = 20 (12M, 8F); HC = 11 (6M, 5F)	Audio recordings of routine medical assessment	Acoustic features	SVM,LR,NB, RF, MLP, and DT	ACC=63.97% to 91.76% (with CV)
		Chakraborty et al.[34]	2018	SCZ = 52 (25M, 27F); HC = 26 (12M, 14F)	Audio recordings of semi-structured clinical interview	Acoustic/prosodic features	Linear SVM	ACC = 79.49% (with CV)
		Tahir et al.[143]	2016	SCZ = 8 (4M, 4F); HC = 7 (2M, 5F)	Audio recordings of semi-structured clinical interview	Conversational features	SVM and SVR	ACC = 86.0 to 93.0% (with CV)
		Martínez-Sánchez et al.[282]	2015	SCZ = 45 (32M, 13F); HC = 35 (22M, 13F)	Audio recordings of semi-structured clinical interview	Acoustic/prosodic features	LDA	ACC = 87.5% (with CV)
		Rapcan et al.[142]	2010	SCZ = 39 (32M, 13F); HC = 18 (22M, 13F)	Digitally recordings of reading aloud a text passage	Acoustic features	LDA	ACC = 79.4% (with CV)
	Facial expressions	Tron et al.[37]	2016	SCZ = 34; HC = 33	Video recordings of structured interview	Facial cluster features	SVM	AUC = 0.80 to 0.85 (with CV)
		Tron et al.[184]	2015	SCZ = 34; HC = 33	Video recordings of structured interview	Facial AUs	SVM	AUC = 0.80 (with CV)
	Body movement	Chakraborty et al.[39]	2017	SCZ = 46 (23M, 23F); HC = 23 (11M, 12F)	Kinect recording of semi-structured clinical interview	linear speed and acceleration of body joints	MLP	ACC=86.76%; AUC=0.913 (with CV)
	Multiple modalities	Xu et al.[117]	2019	SCZ = 43 (21M, 22F); DP = 45 (23F,22F) HC = 41 (23M, 18F)	Audio recordings of semi-structured clinical interview	Verbal: linguistic features Audio: acoustic/prosodic features	Ensemble classifier	SCZ vs. HC: ACC=76.19% DP vs. HC: ACC=72.09% SCZ vs. DP: ACC=69.32% (all with CV)

TABLE B.1: Overview of related behavioral data-driven studies on diagnosis of schizophrenia and depression.

Task	Aspect	Reference	Year	Dataset	Data Source	Features	Classifier	Performance (ACC/BAC/AUC)
DP vs. HC	Verbal (spoken language)	Qureshi et al.[141]	2020	DAIC dataset: 138 samples split into 5 classes	interview transcriptions of human-agent interaction	Sentence encoding network	LSTM	ACC=66%, F1=0.60 (5-class classification, with CV)
	Non-verbal (acoustic/ prosodic/ conversational)	Scherer et al.[283]	2013	DAIC dataset: DP-severe = 18; DP-low = 18	Audio and video recordings of human-computer interaction experiment	Acoustic features	SVM	F1-avg=0.738; ACC=75%
		Espinola et al.[173]	2021	DP = 22 (17M, 5F); HC = 11(6M, 5F)	Audio recordings of medical evaluation interview	Acoustic features	Various Classifiers	ACC=60% to 89.14%
		Low et al.[225]	2011	DP = 68 (19M, 49F); HC = 71 (27M, 44F) (between 14 and 18 years old)	Video recordings of family interation	Acoustic features	GMM-SVM	ACC = 67-87% (with CV)
		Ooi et al.[223]	2013	DP (at risk) = 15 (6M, 9F); HC (no risk) = 15 (6M, 9F) (all adolescent, age 12 to 13 years)	Audio recordings of child-adult interaction	Acoustic Features	GMM- Bayesian classifier	ACC = 73% (with CV)
		Huang et al.[157]	2020	SH2-FS dataset: Train: DP = 97; HC = 364 Test: DF=23; HC=105 DAIC dataset: Train: 107 Test: 35	Audio recordings of free speech in naturalistic environments	Acoustic Features	CNN	DAIC dataset: BAC=91%, F1-avg=0.915; SH2-FS: BAC=73%, F1-avg=0.625
		Sanchez et al. [226]	2011	DP = 16; HC = 16	Audio recordings of structured clinical interview	Prosodic and spectral features	SVM	ACC = 81.3% (with CV)
		Taguchi T[227]	2015	DP = 36 (22M, 14F); HC = 36(16M, 20F)	Audio recordings of reading out numbers and verbal fluency task	Second dimension of MFCC	SDA	ACC = 81.9% (no CV)

TABLE B.1: Overview of related behavioral data-driven studies on diagnosis of schizophrenia and depression.

Task	Aspect	Reference	Year	Dataset	Data Source	Features	Classifier	Performance (ACC/BAC/AUC)
DP vs. HC	Facial expressions	Dibeklioğlu H, et al.[189]	2018	Pitt dataset: DP-severe = 58; DP-mild = 35; DP-low =37 (multi-session data)	Audio and video recordings of HRSD clinical interview.	Facial movement	SDAE	ACC=72.6% (with CV)
		Cohn et al.[186]	2009	Pitt dataset: DP-severe = 66; DP-low =41 (multi-session data)	Audio and video recordings of HRSD clinical interview.	Facial movement	SVM	ACC=79% (with CV)
	Body movement	Horigome et al.[53]	2020	DP = 17 (8M, 9F); BD = 14 (7M,7F); HC = 16(6M, 10F)	Kinect RGB+D recording of semi-structured clinical interview	Body movement featrues	SVM-rbf	ACC=72% (with CV)
	Multiple modalities	Joshi et al.[54]	2013	BlackDog dataset: DP = 30; HC = 30	Audio and video recordings of clinical interview	Audio: acoustic features; Video: Spatio-temporal descriptor	SVM	Audio: ACC = 78.3 to 83.3%; Video: ACC = 78.8 to 81.7%; Audio + Video = 66.7 to 91.7% (no CV and obtained by fine tuning the parameters)
		Yang et al.[166]	2016	DAIC dataset: DP-severe = 7; DP-low = 28	Audio and video recordings of human-agent interaction	Facial expression, facial movement features	SVM	Test Set: F1=0.724; BAC=75.4%;
		Alghowinem et al.[163]	2015	BlackDog dataset: DP = 30; HC = 30(30M, 30F) Pitt dataset: DP-severe = 19; HC(Symptom-free) = 19(14M, 24F) DAIC dataset: DP-severe = 16; DP-low = 16 (9M, 23F)	BlackDog dataset: Audio and video recordings of open-ended questions interview. Pitt dataset: Audio and video recordings of HRSD clinical interview. DAIC dataset: Audio/Video recordings of human-agent interaction	Eye activity and head pose	SVM	BlackDog: ACC=76.7% Pitt: ACC=94.7% DAIC: ACC=68.8% All three combined: ACC=73.1 (with CV)
		Dibeklioğlu et al.[224]	2015	Pitt dataset: DP-severe = 58; DP-low = 37 (multi-session data)	Audio and video recordings of HRSD clinical interview.	Facial movement, head movement, and Audio	LR	Facial: ACC=81.44% Head: ACC=79.59% Audio: ACC=69.73% All three combined: ACC=88.93% (with CV)

TABLE B.1: Overview of related behavioral data-driven studies on diagnosis of schizophrenia and depression.

Task	Aspect	Reference	Year	Dataset	Data Source	Features	Classifier	Performance (ACC/BAC/AUC)
DP vs. HC	Multiple modalities	Valstar et al.[161]	2016	DAIC-WOZ (NA)	Audio/Video recordings of human-agent interaction	Audio: acoustic features; Video: facial and eye expressions	SVM	Test data: Audio: F1-avg = 0.50; Video: F1-avg = 0.72; Audio + Video: F1-avg = 0.72
		Alghowinem et al.[62]	2018	BlackDog dataset: DP = 30; HC = 30	Audio and video recordings of open ended questions interview.	Audio: acoustic and conversational features; Video: Eye and heald movement features	SVM	Audio: BAC = 81.7%; Video: BAC = 63.3 to 78.3%; Audio + Video = 63.3 to 86.7% (with CV)
		Joshi et al.[169]	2013	BlackDog dataset: DP = 30; HC = 30	Audio and video recordings of open ended questions interview.	Facial and upper body movement	SVM	Facial: ACC=71%, F1 = 0.73; Body movement: ACC=77%, F1=0.8 (with CV)
DP vs. SCZ	Verbal (spoken language)	Lott et al.[57]	2002	SCZ = 47 (33M, 14F); BD = 29 (17M, 12F); DP = 23 (10M, 13F)	Manual transcriptions of the structured clinical interview	Linguistic features	LDA	SCZ: Recall=74.1%; BD: Recall=63.3%; DP: Recall=82.4%; Three-class classification: ACC=72.7% (with CV)
	Non-verbal (acoustic / prosodic / conversational)	Kliper et al.[58]	2010	SCZ = 22; DP = 20 HC = 20	NA	Acoustic features	Linear Classifier	SCZ vs. HC: ACC=73.8 to 81.0% DP vs. HC: ACC=69.0 to 87.1% SCZ vs. DP: ACC=52.0 to 76.7% (with CV)

Abbreviation: SCZ=Patient with schizophrenia, HC=Healthy control, DP=Depression patient, BD=Bipolar disorder, TLC=the Assessment of Thought, Language and Communication, BPRS=the Brief Psychiatric Rating Scale, AUs=Action units, LDA=Linear Discriminant Analysis, LSA=Latent Semantic Analysis, LR=Logistic Regression, RF=Random Forest, SVM=Support Vector Machines, XGB=XGBoost, KNN=K-Nearest Neighbors, SVR=Support Vector Regression, LIWC=Linguistic Inquiry and Word Count, ASR=Automatic speech recognition, MLP=Multilayer Perceptron, HDRS=Hamilton Depression Rating Scale, BDI=Beck's Depression Inventory, SDA=Stepwise discriminant analysis, HRSD=Hamilton Rating Scale for Depression, SDAE=Stacked denoising auto-encoders, ACC=Accuracy, BAC=Balanced accuracy, AUC=Area under ROC curve, F1=F1 score, MAE=Mean absolute error, RMSE=Root mean square error, CV=Cross-validation, DAIC=Distress Analysis Interview Corpus.

TABLE B.2: Overview of related behavioral data-driven studies on predicting the severity of negative, cognitive, and general psychiatric symptoms of schizophrenia and depression.

Sample	Aspect	Reference	Year	Dataset	Data Source	Features	Classifier	Assessment	Performance (ACC/Pearson's R/MAE/RMSE)
SCZ	Verbal (spoken language)	Xu et al. [30]	2018	SCZ = 50 (25M, 25F); HC = 25 (11M, 14F)	Kaldi's transcriptions of semi-structured clinical interview	LIWC, DICTION7.0, Doc2Vec	Ensemble classifier	NSA-16	ACC: NSA-Prolonged time to respond=82.69%; NSA-Restricted speech quantity=82.69%; NSA-Impoverished speech content=80.77%; NSA-Emotion reduced range=61.54%; NSA-Reduced modeulation of intensity=78.85%; NSA-Reduced expressive gestures=84.62%
	Non-verbal (acoustic/ prosodic/ conversational)	Cohen et al.[56]	2020	121 patients (SCZ=76, DP=18, BD=20, others=7)	Audio recordings of picture task and free speech for daily routines	Acoustic features	LASSO	BPRS and SANS	SANS-Alogia: ACC=85% to 87%, SEN=0.60 to 0.70, SPE=0.92 to 0.93; SANS-Blunted Vocal Affect: ACC=89% to 96%, SEN=0.66 to 0.82, SEN=0.92 to 0.99
		Chakraborty et al.[34]	2018	SCZ = 52 (25M, 27F); HC = 26 (12M, 14F)	Audio recordings of semi-structured clinical interview	Acoustic/prosodic features	KNN, SVM, and AdaBoost	NSA-16	ACC: NSA-Prolonged time to respond=82.69%; NSA-Restricted speech quantity=82.69%; NSA-Impoverished speech content=80.77%; NSA-Emotion reduced range=61.54%; NSA-Reduced modulation of intensity=78.85%; NSA-Reduced expressive gestures=84.62%
		Wörtwein et al.[271]	2017	20 psychiatric individuals with SCZ, DP, or mania	Audio recordings of semi-structured clinical interview	Acoustic features	SVR	BPRS-24	Pearson' R/MAE: BPRS-Elevated mood=0.68/0.80; BPRS-Grandiosity=0.80/0.46; BPRS-Excitement=0.72/0.69; BPRS-Motor hyperactivity=0.67/0.84; BPRS-Total=0.49/10.35
	Facial expressions	Bishay et al.[185]	2019	91 out-patients	Video recordings of body psychotherapy of schizophrenia	Facial expressions	DNN-GMM	PANSS	PANSS-NEG: MAE = 3.35 and RMSE=4.27
		Tron et al.[37]	2016	SCZ = 34; HC = 33	Video recordings of structured interview	Facial cluster features	Ridge regression	PANSS	Pearson's R between predictions and scales: PANSS-Blunted affect=0.431

TABLE B.2: Overview of related behavioral data-driven studies on predicting the severity of negative, cognitive, and general psychiatric symptoms of schizophrenia and depression.

Sample	Aspect	Reference	Year	Dataset	Data Source	Features	Classifier	Assessment	Performance (ACC/Pearson's R/MAE/RMSE)
SCZ	Facial expressions	Tron et al.[184]	2015	SCZ = 34; HC = 33	Video recordings of structured interview	Facial AUs	Ridge regression	PANSS	Pearson's R between predictions and scales: PANSS-Blunted affect=0.530; PANSS-Emotional withdrawal=0.510; PANSS-Difficulty in abstract thinking =0.369; PANSS-Stereotyped thinking =0.369; PANSS-Passive withdrawal=0.368
	Body movement	Chakraborty et al.[39]	2017	SCZ = 46 (23M, 23F); HC = 23 (11M, 12F)	Kinect recording of semi-structured clinical interview	Linear Speed and Accelaration of Body Joints	KNN and SVM	NSA-16	ACC: NSA-Prolonged time to respond=60.87%; NSA-Restricted speech quantity=82.69%%; NSA-Impoverished speech content=67.39%; NSA-Emotion reduced range=61.54%; NSA-Reduced modeulation of intensity=63.04%
	Multiple modalities	Xu et al.[117]	2019	SCZ = 43 (21M, 22F); DP = 45 (23F,22F) HC = 41 (23M, 18F)	Audio recordings of semi-structured clinical interview	Verbal: linguistic features Audio: acoustic/prosodic features	Ensemble classifier	NSA-16	ACC: NSA-Restricted speech quantity=90.5%; NSA-Reduced Display=71.4%; NSA-Reduced expressive gestures=64.3%
SCZ & DP	Non-verbal (acoustic/ prosodic/ conversational)	Cohen AS, et al. [56]	2020	121 patients (SCZ=76, DP=18, BD=20, others=7)	Audio recordings of picture task and free speech for daily routines	Acoustic features	LASSO	BPRS and SANS	SANS-Alogia: ACC=85% to 87%, SEN=0.60 to 0.70, SPE=0.92 to 0.93; SANS-Blunted Vocal Affect: ACC=89% to 96%, SEN=0.66 to 0.82, SEN=0.92 to 0.99
		Wörtwein et al.[271]	2017	20 psychiatric individuals with SCZ, DP, or mania	Audio recordings of semi-structured clinical interview	Acoustic features	SVR	BPRS-24	Pearson' R/MAE: BPRS-Elevated mood=0.68/0.80; BPRS-Grandiosity=0.80/0.46; BPRS-Excitement=0.72/0.69; BPRS-Motor hyperactivity=0.67/0.84; BPRS-Total=0.49/10.35
	Multiple Modalities	Xu S, et al. [117]	2019	SCZ = 43 (21M, 22F); DP = 45 (23F,22F) HC = 41 (23M, 18F)	Audio recordings of semi-structured clinical interview	Verbal: linguistic features Audio: acoustic/prosodic features	Voting Classifier	NSA-16	ACC: NSA-Impoverished speech content=70.5%; NSA-Emotion reduced range=70.5%; NSA-Reduced sexual interest=68.2%; NSA-Reduced expressive gestures=70.5%

Abbreviation: SCZ=Patient with schizophrenia, HC=Healthy control, DP=Depression patient, BD=Bipolar disorder, NSA-16=The 16-item Negative Symptom Assessment, BPRS=The Brief Psychiatric Rating Scale, PANSS=The Positive and Negative Syndrome Scale, AUs=Action units, LR=Logistic Regression, RF=Random Forest, SVM=Support Vector Machines, KNN=K-Nearest Neighbors, SVR=Support Vector Regression, ACC=accuracy, BAC=Balanced accuracy, AUC=Area under ROC curve, F1=F1 score, MAE=Mean absolute error, RMSE=Root mean square error, CV=Cross-validation.

List of Author's Awards, Patents, and Publications[1]

Awards

- **Best Paper Runner-Up**, "Automated Lexical Analysis of Interviews with Schizophrenic Patients, International Workshop on Spoken Dialogue Systems Technology (IWSDS 2018)".

Journal Articles

- **Shihao Xu**, Zixu Yang, Debsubhra Chakraborty, Yi Han Victoria Chua, Tolomeo Serenella, Stefan Winkler, Michel Birnbaum, Bhing-Leet Tan, Jimmy Lee Chee Keong, and Justin Dauwels, "Machine Learning Methods to Predict Negative, Cognitive, and General Psychiatric Symptoms of Schizophrenia and Depression Patients from Behavioral Cues," *npj Schizophrenia*, submitted.

Conference Proceedings

- **Shihao Xu**, Zixu Yang, Debsubhra Chakraborty, Yi Han Victoria Chua, Justin Dauwels, Daniel Thalmann, Nadia Magnenat Thalmann, Bhing-Leet Tan, Jimmy Lee Chee Keong, "Automated Verbal and Non-verbal Speech Analysis of Interviews of Individuals with Schizophrenia and Depression," in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019.*

[1]The superscript * indicates joint first authors

- **Shihao Xu**, Zixu Yang, Debsubhra Chakraborty, Yasir Tahir, Tomasz Maszczyk, Yi Han Victoria Chua, Justin Dauwels, Daniel Thalmann, Nadia Magnenat Thalmann, Bhing-Leet Tan, Jimmy Lee Chee Keong, "Automatic Verbal Analysis of Interviews with Schizophrenic Patients," in *IEEE 23rd International Conference on Digital Signal Processing (DSP), 2018.*

- **Shihao Xu**, Zixu Yang, Debsubhra Chakraborty, Yasir Tahir, Tomasz Maszczyk, Yi Han Victoria Chua, Justin Dauwels, Daniel Thalmann, Nadia Magnenat Thalmann, Bhing-Leet Tan, Jimmy Lee Chee Keong, "Automated Lexical Analysis of Interviews with Individuals with Schizophrenia," in *9th International Workshop on Spoken Dialogue System Technology (IWSDS), 2018.*

- Debsubhra Chakraborty, **Shihao Xu**, Zixu Yang, Yi Han Victoria Chua, Yasir Tahir, Justin Dauwels, Nadia Thalmann Magnenat, Bhing-Leet Tan, Jimmy Lee Chee Keong, "Prediction of Negative Symptoms of Schizophrenia from Objective Linguistic, Acoustic and Non-verbal Conversational Cues," in *International Conference on Cyberworlds, 2018.*

Bibliography

[1] O. Yildiz and A. Arslan, "Automated Auscultative Diagnosis System for Evaluation of Phonocardiogram Signals Associated with Heart Murmur Diseases," *Gazi University Journal of Science*, vol. 31, no. 1, pp. 112–124, Mar. 2018, number: 1. [Online]. Available: https://dergipark.org.tr/en/pub/gujs/issue/35772/339978 xii, 28

[2] G. Lam, H. Dongyan, and W. Lin, "Context-aware Deep Learning for Multi-modal Depression Detection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 3946–3950. xii, 29

[3] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 10, pp. 1683–1699, 2007. xii, 31

[4] M. Bishay, P. Palasek, S. Priebe, and I. Patras, "SchiNet: Automatic Estimation of Symptoms of Schizophrenia from Facial Behaviour Analysis," *IEEE Trans. Affective Comput.*, pp. 1–1, 2019. xii, 32

[5] M. Nasir, B. Baucom, C. Bryan, S. Narayanan, and P. Georgiou, "Modeling Vocal Entrainment in Conversational Speech using Deep Unsupervised Learning," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020, conference Name: IEEE Transactions on Affective Computing. xii, 9, 36, 37

[6] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, Abbasi *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017," *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, Nov. 2018. 1

[7] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition ed. American Psychiatric Association, May 2013. 1, 2, 12

[8] T. M. Laursen, T. Munk-Olsen, and M. Vestergaard, "Life expectancy and cardiovascular mortality in persons with schizophrenia," *Current Opinion in Psychiatry*, vol. 25, no. 2, pp. 83–88, Mar. 2012. 1

[9] M. Savill, S. Orfanos, U. Reininghaus, T. Wykes, R. Bentall, and S. Priebe, "3. The relationship between experiential deficits of negative symptoms and subjective quality of life in schizophrenia," *Schizophrenia Research*, vol. 176, no. 2, pp. 387–391, Oct. 2016. 1

[10] G. P. Strauss and A. S. Cohen, "A Transdiagnostic Review of Negative Symptom Phenomenology and Etiology," *Schizophrenia Bulletin*, vol. 43, no. 4, pp. 712–719, Jul. 2017. 1, 2

[11] S. B. Guessoum, Y. Le Strat, C. Dubertret, and J. Mallet, "A transnosographic approach of negative symptoms pathophysiology in schizophrenia and depressive disorders," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 99, p. 109862, Apr. 2020. 1, 10

[12] P. L. Rock, J. P. Roiser, W. J. Riedel, and A. D. Blackwell, "Cognitive impairment in depression: a systematic review and meta-analysis," *Psychological Medicine*, vol. 44, no. 10, pp. 2029–2040, Jul. 2014, publisher: Cambridge University Press. 1

[13] J. Lim, S.-A. Lee, M. Lam, A. Rapisarda, M. Kraus, R. S. E. Keefe, and J. Lee, "The relationship between negative symptom subdomains and cognition," *Psychological Medicine*, vol. 46, no. 10, pp. 2169–2177, Jul. 2016, publisher: Cambridge University Press. 1, 122

[14] M.-P. Austin, P. Mitchell, and G. M. Goodwin, "Cognitive deficits in depression: Possible implications for functional neuropathology," *Br J Psychiatry*, vol. 178, no. 3, pp. 200–206, Mar. 2001. 1

[15] Hammar and G. Årdal, "Cognitive Functioning in Major Depression – A Summary," *Frontiers in Human Neuroscience*, vol. 3, 2009. 1

[16] J. Ventura, K. L. Subotnik, M. J. Gitlin, D. Gretchen-Doorly, A. Ered, K. F. Villa, G. S. Hellemann, and K. H. Nuechterlein, "Negative symptoms and functioning during the first year after a recent onset of schizophrenia and 8years later," *Schizophrenia Research*, vol. 161, no. 2, pp. 407–413, Feb. 2015. 1

[17] C. U. Correll and N. R. Schooler, "Negative Symptoms in Schizophrenia: A Review and Clinical Guide for Recognition, Assessment, and Treatment," *Neuropsychiatr Dis Treat*, vol. 16, pp. 519–534, Feb. 2020. 2, 127

[18] A. S. Cohen, E. Schwartz, T. Le, T. Cowan, C. Cox, R. Tucker, P. Foltz, T. B. Holmlund, and B. Elvevåg, "Validating digital phenotyping technologies for clinical use: the critical importance of "resolution"," *World Psychiatry*, vol. 19, no. 1, pp. 114–115, 2020, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wps.20703. 2

[19] U. Granziol, A. Spoto, and G. Vidotto, "The assessment of nonverbal behavior in schizophrenia through the Formal Psychological Assessment," *International Journal of Methods in Psychiatric Research*, vol. 27, no. 1, p. e1595, 2018, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mpr.1595. 2

[20] L. D. Alphs, A. Summerfelt, H. Lann, and R. J. Muller, "The negative symptom assessment: a new instrument to assess negative symptoms of schizophrenia." *Psychopharmacol Bull*, vol. 25, no. 2, pp. 159–163, 1989, place: United States. 2, 33, 45

[21] B. Kirkpatrick, W. S. Fenton, W. T. Carpenter, Jr., and S. R. Marder, "The NIMH-MATRICS Consensus Statement on Negative Symptoms," *Schizophrenia Bulletin*, vol. 32, no. 2, pp. 214–219, Apr. 2006. 2

[22] A. S. Cohen, S. C. Morrison, L. A. Brown, and K. S. Minor, "Towards a cognitive resource limitations model of diminished expression in schizotypy." *Journal of Abnormal Psychology*, vol. 121, no. 1, pp. 109–118, 2012. 2

[23] A. S. Cohen, J. E. McGovern, T. J. Dinzeo, and M. A. Covington, "Speech deficits in serious mental illness: A cognitive resource issue?" *Schizophrenia Research*, vol. 160, no. 1, pp. 173–179, Dec. 2014. 3

[24] J.-P. Onnela and S. L. Rauch, "Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health," *Neuropsychopharmacology*, vol. 41, no. 7, pp. 1691–1696, Jun. 2016, number: 7 Publisher: Nature Publishing Group. 3

[25] L. Marzano, A. Bardill, B. Fields, K. Herd, D. Veale, N. Grey, and P. Moran, "The application of mHealth to mental health: opportunities and challenges," *The Lancet Psychiatry*, vol. 2, no. 10, pp. 942–948, Oct. 2015. 3

[26] T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, D. S. Pine, K. Quinn, C. Sanislow, and P. Wang, "Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders," *AJP*, vol. 167, no. 7, pp. 748–751, Jul. 2010, publisher: American Psychiatric Publishing. 3

[27] A. S. Cohen, G. M. Najolia, Y. Kim, and T. J. Dinzeo, "On the boundaries of blunt affect/alogia across severe mental illness: Implications for Research Domain Criteria," *Schizophrenia Research*, vol. 140, no. 1-3, pp. 41–45, Sep. 2012. 3, 26, 121

[28] J. Torous, M. V. Kiang, J. Lorme, and J.-P. Onnela, "New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research," *JMIR Mental Health*, vol. 3, no. 2, p. e5165, May 2016, company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada. 3

[29] J. N. de Boer, S. G. Brederoo, A. E. Voppel, and I. E. C. Sommer, "Anomalies in language as a biomarker for schizophrenia," *Current Opinion in Psychiatry*, vol. 33, no. 3, pp. 212–218, May 2020. 3

[30] S. Xu, Z. Yang, D. Chakraborty, Y. Tahir, T. Maszczyk, V. Y. H. Chua, J. Dauwels, D. Thalmann, N. M. Thalmann, B. Tan, and J. L. C. Keong, "Automatic Verbal Analysis of Interviews with Schizophrenic Patients," in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, Nov. 2018, pp. 1–5. 39, 133, 138

[31] E. S. Kayi, M. Diab, L. Pauselli, M. Compton, and G. Coppersmith, "Predictive Linguistic Features of Schizophrenia," *arXiv:1810.09377 [cs]*, Oct. 2018, arXiv: 1810.09377. 3

[32] A. Parola, A. Simonsen, V. Bliksted, and R. Fusaroli, "Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis," *Schizophrenia Research*, vol. 216, pp. 24–40, Feb. 2020. 3, 24, 38, 75

[33] 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, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lio2.354. 27

[34] D. Chakraborty, Z. Yang, Y. Tahir, T. Maszczyk, J. Dauwels, N. Thalmann, J. Zheng, Y. Maniam, N. Amirah, B. L. Tan, and J. Lee, "Prediction of Negative Symptoms of Schizophrenia from Emotion Related Low-Level Speech Signals," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 6024–6028. 3, 26, 39, 75, 134, 138

[35] Y. Tahir, Z. Yang, D. Chakraborty, N. Thalmann, D. Thalmann, Y. Maniam, N. A. b. A. Rashid, B.-L. Tan, J. L. C. Keong, and J. Dauwels, "Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia," *PLOS ONE*, vol. 14, no. 4, p. e0214314, Apr. 2019, publisher: Public Library of Science. 3, 39, 56, 119

[36] A. S. Cohen, E. Schwartz, T. P. Le, T. Cowan, B. Kirkpatrick, I. M. Raugh, and G. P. Strauss, "Digital phenotyping of negative symptoms: the relationship to clinician ratings," *Schizophrenia Bulletin*, vol. 47, no. 1, pp. 44–53, Jan. 2021.

[37] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall, "Facial expressions and flat affect in schizophrenia, automatic analysis from depth camera data," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Feb. 2016, pp. 220–223. 32, 39, 122, 134, 138

[38] H.-H. Tseng, S.-H. Chen, C.-M. Liu, O. Howes, Y.-L. Huang, M. H. Hsieh, C.-C. Liu, J.-C. Shan, Y.-T. Lin, and H.-G. Hwu, "Facial and Prosodic Emotion Recognition Deficits Associate with Specific Clusters of Psychotic Symptoms in Schizophrenia," *PLOS ONE*, vol. 8, no. 6, p. e66571, Jun. 2013. 3

[39] D. Chakraborty, Y. Tahir, Z. Yang, T. Maszczyk, J. Dauwels, D. Thalmann, N. M. Thalmann, B. Tan, and J. Lee, "Assessment and prediction of negative symptoms of schizophrenia from RGB+D movement signals," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, Oct. 2017, pp. 1–6. 3, 35, 39, 119, 134, 139

[40] S. Walther, F. Ramseyer, H. Horn, W. Strik, and W. Tschacher, "Less Structured Movement Patterns Predict Severity of Positive Syndrome, Excitement, and Disorganization," *Schizophrenia Bulletin*, vol. 40, no. 3, pp. 585–591, May 2014. 34

[41] Z. Kupper, F. Ramseyer, H. Hoffmann, S. Kalbermatten, and W. Tschacher, "Video-based quantification of body movement during social interaction indicates the severity of negative symptoms in patients with schizophrenia," *Schizophrenia Research*, vol. 121, no. 1, pp. 90–100, Aug. 2010. 3, 33

[42] A. Arseniev-Koehler, S. Mozgai, and S. Scherer, "What type of happiness are you looking for? - A closer look at detecting mental health from language," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. New Orleans, LA: Association for Computational Linguistics, Jun. 2018, pp. 1–12. 3

[43] J. Zimmermann, T. Brockmeyer, M. Hunn, H. Schauenburg, and M. Wolf, "First-person Pronoun Use in Spoken Language as a Predictor of Future Depressive Symptoms: Preliminary Evidence from a Clinical Sample of Depressed Patients," *Clinical Psychology & Psychotherapy*, vol. 24, no. 2, pp. 384–391, 2017. 23

[44] R. Trifu, B. Nemes, C. Bodea-Haţegan, and D. Cozman, "Linguistic indicators of language in major depressive disorder (MDD). An evidence based research," *Journal of Evidence-Based Psychotherapies*, vol. 17, pp. 105–128, Mar. 2017. 3, 23, 39, 73

[45] 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)*, Dec. 2016, pp. 136–143. 3, 22, 39

[46] M. Yamamoto, A. Takamiya, K. Sawada, M. Yoshimura, M. Kitazawa, K.-c. Liang, T. Fujita, M. Mimura, and T. Kishimoto, "Using speech recognition technology to investigate the association between timing-related speech features and depression severity," *PLoS ONE*, vol. 15, no. 9, p. e0238726, Sep. 2020. 30, 73, 74

[47] M. Neumann, O. Roessler, D. Suendermann-Oeft, and V. Ramanarayanan, "On the Utility of Audiovisual Dialog Technologies and Signal Analytics for Real-time Remote Monitoring of Depression Biomarkers," in *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 47–52. 3, 4

[48] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, "Multimodal assessment of depression from behavioral signals," S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, Eds. New York, NY, USA: Association for Computing Machinery and Morgan & Claypool, 2019, pp. 375–417.

[49] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, Jul. 2015. 3, 27, 38, 39

[50] A. Pampouchidou, P. G. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, and M. Tsiknakis, "Automatic Assessment of Depression Based on Visual Cues: A Systematic Review," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 445–470, Oct. 2019, conference Name: IEEE Transactions on Affective Computing. 3, 4, 33, 120

[51] L. He, D. Jiang, and H. Sahli, "Automatic Depression Analysis Using Dynamic Facial Appearance Descriptor and Dirichlet Process Fisher Encoding," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1476–1486, Jun. 2019. 33

[52] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image and Vision Computing*, vol. 32, no. 10, pp. 641–647, Oct. 2014. 3, 32, 34, 94, 95

[53] T. Horigome, B. Sumali, M. Kitazawa, M. Yoshimura, K.-c. Liang, Y. Tazawa, T. Fujita, M. Mimura, and T. Kishimoto, "Evaluating the severity of depressive symptoms using upper body motion captured by RGB-depth sensors and machine learning in a clinical interview setting: A preliminary study," *Comprehensive Psychiatry*, vol. 98, p. 152169, Apr. 2020. 3, 34, 39, 136

[54] J. Joshi, R. Goecke, G. Parker, and M. Breakspear, "Can body expressions contribute to automatic depression analysis?" in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Apr. 2013, pp. 1–7. 3, 34, 39, 120, 136

[55] A. S. Cohen, T. Cowan, T. P. Le, E. K. Schwartz, B. Kirkpatrick, I. M. Raugh, H. C. Chapman, and G. P. Strauss, "Ambulatory digital phenotyping of blunted affect and alogia using objective facial and vocal analysis: Proof of concept," *Schizophrenia Research*, vol. 220, pp. 141–146, Jun. 2020. 3, 26, 32, 39, 95, 121

[56] A. S. Cohen, C. R. Cox, T. P. Le, T. Cowan, M. D. Masucci, G. P. Strauss, and B. Kirkpatrick, "Using machine learning of computerized vocal expression to measure blunted vocal affect and alogia," *npj Schizophrenia*, vol. 6, no. 1, pp. 1–9, Sep. 2020, number: 1 Publisher: Nature Publishing Group. 3, 26, 75, 138, 139

[57] P. R. Lott, S. Guggenbühl, A. Schneeberger, A. E. Pulver, and H. H. Stassen, "Linguistic Analysis of the Speech Output of Schizophrenic, Bipolar, and Depressive Patients," *PSP*, vol. 35, no. 4, pp. 220–227, 2002. 3, 23, 75, 120, 137

[58] R. Kliper, Y. Vaizman, D. Weinshall, and S. Portuguese, "Evidence for depression and schizophrenia in speech prosody," 2010, pp. 85–88. 3, 39, 120, 134, 137

[59] M. J. Kas, B. Penninx, B. Sommer, A. Serretti, C. Arango, and H. Marston, "A quantitative approach to neuropsychiatry: The why and the how," *Neuroscience & Biobehavioral Reviews*, vol. 97, pp. 3–9, Feb. 2019. 4, 120

[60] B. N. Cuthbert, "The role of RDoC in future classification of mental disorders," *Dialogues Clin Neurosci*, vol. 22, no. 1, pp. 81–85, Mar. 2020. 4, 75

[61] S. M. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, and G. Parker, "Interpretation of Depression Detection Models via Feature Selection Methods," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020, conference Name: IEEE Transactions on Affective Computing. 4, 28

[62] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, Oct. 2018, conference Name: IEEE Transactions on Affective Computing. 4, 39, 137

[63] J. Torous, J.-P. Onnela, and M. Keshavan, "New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices," *Transl Psychiatry*, vol. 7, no. 3, pp. e1053–e1053, Mar. 2017. 4

[64] J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli, Eds., *Social Signal Processing*. Cambridge: Cambridge University Press, 2017. 4, 5, 9, 101

[65] A. Vinciarelli, A. Esposito, E. André, F. Bonin, M. Chetouani, J. F. Cohn, M. Cristani, F. Fuhrmann, E. Gilmartin, Z. Hammal, D. Heylen, R. Kaiser, M. Koutsombogera, A. Potamianos, S. Renals, G. Riccardi, and A. A. Salah, "Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions," *Cogn Comput*, vol. 7, no. 4, pp. 397–413, Aug. 2015. 4

[66] J. De Vaus, M. J. Hornsey, P. Kuppens, and B. Bastian, "Exploring the East-West Divide in Prevalence of Affective Disorder: A Case for Cultural Differences in Coping With Negative Emotion," *Pers Soc Psychol Rev*, vol. 22, no. 3, pp. 285–304, Aug. 2018, publisher: SAGE Publications Inc. 5

[67] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009. 5

[68] U. Eco, *A Theory of Semiotics*. Indiana University Press, 1979, google-Books-ID: BoXO4ItsuaMC. 5

[69] P. Ekman and W. V. Friesen, *Manual for the facial action coding system*. Consulting Psychologists Press, 1978. 6

[70] M. Argyle and M. Cook, *Gaze and mutual gaze*. Cambridge, Eng. ; New York: Cambridge University Press, 1976. 6

[71] W. S. Condon and W. D. Ogston, "A segmentation of behavior," *Journal of Psychiatric Research*, vol. 5, no. 3, pp. 221–235, 1967, place: Netherlands Publisher: Elsevier Science. 6

[72] D. Efron, *Gesture and environment*, ser. Gesture and environment. Oxford, England: King'S Crown Press, 1941, pages: x, 184. 6

[73] E. S. Klima and U. Bellugi, *The Signs of Language*. Harvard University Press, 1979, google-Books-ID: WeBOn6N8PJ8C. 6

[74] R. Barocas and P. Karoly, "Effects of physical appearance on social responsiveness," *Psychol Rep*, vol. 31, no. 2, pp. 495–500, Oct. 1972. 6

[75] H. T. Reis, J. Nezlek, and L. Wheeler, "Physical attractiveness in social interaction," *Journal of Personality and Social Psychology*, vol. 38, no. 4, pp. 604–617, 1980, place: US Publisher: American Psychological Association. 6

[76] J. Streeck, "Gesture as communication I: Its coordination with gaze and speech," *null*, vol. 60, no. 4, pp. 275–299, Dec. 1993, publisher: Routledge. 7

[77] A. Mehrabian, "Significance of posture and position in the communication of attitude and status relationships," *Psychological Bulletin*, vol. 71, no. 5, pp. 359–372, 1969, place: US Publisher: American Psychological Association. 7

[78] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychological Bulletin*, pp. 256–274, 1992. 8

[79] P. Ekman, "Facial action coding system (FACS)," *A human face*, 2002. 8, 31

[80] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, Jan. 2003. 8

[81] J. L. Tracy, R. W. Robins, and R. A. Schriber, "Development of a FACS-verified set of basic and self-conscious emotion expressions," *Emotion*, vol. 9, no. 4, pp. 554–559, Aug. 2009. 8

[82] M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, Aug. 2000. 8

[83] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans. Affective Comput.*, pp. 1–1, 2020, arXiv: 1804.08348. 8

[84] D. Crystal, *Prosodic Systems and Intonation in English*. CUP Archive, 1969, google-Books-ID: aS45AAAAIAAJ. 8

[85] N. Ferguson, "Simultaneous speech, interruptions and dominance," *British Journal of Social and Clinical Psychology*, vol. 16, no. 4, pp. 295–302, 1977, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8260.1977.tb00235.x. 8

[86] A. M. Goberman, "Correlation between acoustic speech characteristics and non-speech motor performance in Parkinson Disease," *Med Sci Monit*, vol. 11, no. 3, pp. CR109–CR116, Mar. 2005, publisher: International Scientific Information, Inc. 9

[87] Y. Tahir, D. Chakraborty, T. Maszczyk, S. Dauwels, J. Dauwels, N. Thalmann, and D. Thalmann, "Real-time sociometrics from audio-visual features for two-person dialogs," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*. Singapore, Singapore: IEEE, Jul. 2015, pp. 823–827. 9, 61

[88] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," *arXiv:1003.4083 [cs]*, Mar. 2010, arXiv: 1003.4083. 9

[89] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543–565, May 2012. 9

[90] I. Altman, "The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding," 1975, publisher: Brooks/Cole Publishing Company, Monterey, California 93940 ($6. 9

[91] K. Emina and G. Jan, "F0 accommodation and turn competition in overlapping talk," *Journal of Phonetics*, vol. 71, pp. 376–394, Nov. 2018. 9, 35

[92] M. J. Pickering and S. Garrod, "Alignment as the basis for successful communication," *Research on Language and Computation*, vol. 4, no. 2-3, pp. 203–228, 2006. 9

[93] C. De Looze, C. Oertel, S. Rauzy, and N. Campbell, "Measuring dynamics of mimicry by means of prosodic cues in conversational speech," in *ICPhS 2011*, Hong-Kong, China, 2011. 9

[94] S. Kousidis, D. Dorran, C. Mcdonnell, and E. Coyle, "Convergence in Human Dialogues Time Series Analysis of Acoustic Feature," Jun. 2019. 9

[95] J. Burgoon, L. Stern, and L. Dillman, "Interpersonal Adaptation: Dyadic Interaction Patterns," Oct. 1995. 9, 10

[96] H. Giles, "Communication Accommodation Theory," in *The International Encyclopedia of Communication Theory and Philosophy*. American Cancer Society, 2016, pp. 1–7, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118766804.wbiect056. 10, 127

[97] F. R. Bilous and R. M. Krauss, "Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads," *Language & Communication*, vol. 8, no. 3, pp. 183–194, Jan. 1988. 10

[98] R. L. Street, "Speech Convergence and Speech Evaluation in Fact-Finding Interviews," *Human Comm Res*, vol. 11, no. 2, pp. 139–169, Dec. 1984. 10

[99] J. Cappella and S. PLANALP, "Talk and silence sequences in informal conversations III: Interspeaker influence," *Human Communication Research*, vol. 7, pp. 117–132, Mar. 2006. 10

[100] F. J. Bernieri and R. Rosenthal, "Interpersonal coordination: Behavior matching and interactional synchrony." in *Fundamentals of nonverbal behavior.*, ser. Studies in emotion & social interaction. Paris, France: Editions de la Maison des Sciences de l'Homme, 1991, pp. 401–432. 10

[101] J. K. Burgoon, N. E. Dunbar, and H. Giles, "Interaction Coordination and Adaptation," in *Social Signal Processing*, J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli, Eds. Cambridge: Cambridge University Press, 2017, pp. 78–96. 10, 127

[102] C. Demily and N. Franck, "Cognitive remediation: a promising tool for the treatment of schizophrenia," *Expert Review of Neurotherapeutics*, vol. 8, no. 7, pp. 1029–1036, Jul. 2008, publisher: Taylor & Francis _eprint: https://doi.org/10.1586/14737175.8.7.1029. 10

[103] S. Mitra, T. Mahintamani, A. R. Kavoor, and S. H. Nizamie, "Negative symptoms in schizophrenia," *Ind Psychiatry J*, vol. 25, no. 2, pp. 135–144, 2016. 11

[104] World Health Organization, *Mental health atlas 2014*. World Health Organization, 2015. [Online]. Available: https://apps.who.int/iris/handle/10665/178879 11

[105] "The World Health Report 2001: Mental Disorders affect one in four people." [Online]. Available: https://www.who.int/news/item/28-09-2001-the-world-health-report-2001-mental-disorders-affect-one-in-four-people 11

[106] K. J. Steinman, A. B. Shoben, A. E. Dembe, and K. J. Kelleher, "How Long Do Adolescents Wait for Psychiatry Appointments?" *Community Ment Health J*, vol. 51, no. 7, pp. 782–789, Oct. 2015. 11

[107] K. Crittenden, C. Yavorsky, F. Ockun, K. Wolanski, and K. A. Kobak, "Methods for Determining Inter-Rater Reliability of the PANSS: A Review of the Literature: (625992009-001)," American Psychological Association, Tech. Rep., 2009. [Online]. Available: http://doi.apa.org/get-pe-doi.cfm?doi=10.1037/e625992009-001 11

[108] "How Much Does a Psychiatrist Cost Without Insurance?" Sep. 2021. [Online]. Available: https://www.talkspace.com/blog/how-much-does-a-psychiatrist-cost/ 11

[109] D. Susman, "8 Reasons Why People Don't Get Treatment for Mental Illness," *David Susman PhD Resources & Inspiration for Better Mental Health*, 2015. 12

[110] "Thinking big in mental health," *Nat. Med.*, vol. 24, no. 1, p. 1, 2018. 17, 38

[111] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Sci Rep*, vol. 6, no. 1, p. 26094, May 2016, number: 1 Publisher: Nature Publishing Group. 17

[112] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 760–772, Oct. 2009. 17

[113] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC2015," Sep. 2015, accepted: 2015-09-16T13:00:41Z. 18, 56, 58

[114] T. Loughran and B. McDonald, "The Use of Word Lists in Textual Analysis," *Journal of Behavioral Finance*, vol. 16, no. 1, pp. 1–11, Jan. 2015, publisher: Routledge _eprint: https://doi.org/10.1080/15427560.2015.1000335. 18, 56, 58

[115] K. Hong, A. Nenkova, M. E. March, A. P. Parker, R. Verma, and C. G. Kohler, "Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls," *Psychiatry Research*, vol. 225, no. 1, pp. 40–49, Jan. 2015. 18, 39, 133

[116] A. St-Hilaire, A. S. Cohen, and N. M. Docherty, "Emotion word use in the conversational speech of schizophrenia patients," *Cognitive Neuropsychiatry*, vol. 13, no. 4, pp. 343–356, Jul. 2008, publisher: Routledge _eprint: https://doi.org/10.1080/13546800802250560. 18, 19

[117] S. Xu, Z. Yang, D. Chakraborty, Y. H. Victoria Chua, J. Dauwels, D. Thalmann, N. M. Thalmann, B.-L. Tan, and J. L. Chee Keong, "Automated Verbal and Non-verbal Speech Analysis of Interviews of Individuals with Schizophrenia and Depression," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2019, pp. 225–228, iSSN: 1558-4615. 18, 26, 39, 74, 119, 134, 139

[118] K. S. Minor, K. A. Bonfils, L. Luther, R. L. Firmin, M. Kukla, V. R. MacLain, B. Buck, P. H. Lysaker, and M. P. Salyers, "Lexical analysis in schizophrenia: How emotion and social word use informs our understanding of clinical presentation," *Journal of Psychiatric Research*, vol. 64, pp. 74–78, May 2015. 18, 19, 74

[119] M. Bell, R. Milstein, J. Beam-Goulet, P. Lysaker, and D. Cicchetti, "The Positive and Negative Syndrome Scale and the Brief Psychiatric Rating Scale: Reliability, Comparability, and Predictive Validity," *The Journal of Nervous and Mental Disease*, vol. 180, no. 11, pp. 723–728, Nov. 1992. 19, 45, 122

[120] S. R. Kay, A. Fiszbein, and L. A. Opler, "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia," *Schizophr Bull*, vol. 13, no. 2, pp. 261–276, Jan. 1987. 19, 45

[121] A. Parola, I. Gabbatore, L. Berardinelli, R. Salvini, and F. M. Bosco, "Multimodal assessment of communicative-pragmatic features in schizophrenia: a machine learning approach," *npj Schizophr*, vol. 7, no. 1, pp. 1–9, May 2021, number: 1 Publisher: Nature Publishing Group. 20, 38, 39

[122] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997. 20

[123] A. K. McCallum, "Mallet: A machine learning for language toolkit," *http://mallet. cs. umass. edu*, 2002. 20, 56, 59

[124] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," *arXiv:1405.4053 [cs]*, May 2014, arXiv: 1405.4053. 20, 56, 59

[125] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 20

[126] C. M. Corcoran, F. Carrillo, D. Fernández-Slezak, G. Bedi, C. Klim, D. C. Javitt, C. E. Bearden, and G. A. Cecchi, "Prediction of psychosis across protocols and risk cohorts using automated language analysis," *World Psychiatry*, vol. 17, no. 1, pp. 67–75, 2018. 21, 39, 133

[127] K. Bar, V. Zilberstein, I. Ziv, H. Baram, N. Dershowitz, S. Itzikowitz, and E. Vadim Harel, "Semantic Characteristics of Schizophrenic Speech," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical*

Psychology. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 84–93. 22, 39, 133

[128] S. X. Tang, R. Kriz, S. Cho, S. J. Park, J. Harowitz, R. E. Gur, M. T. Bhati, D. H. Wolf, J. Sedoc, and M. Y. Liberman, "Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders," *npj Schizophrenia*, vol. 7, no. 1, pp. 1–8, May 2021, number: 1 Publisher: Nature Publishing Group. 22, 39, 133

[129] C. Howes, M. Purver, and R. McCabe, "Using Conversation Topics for Predicting Therapy Outcomes in Schizophrenia," *Biomed Inform Insights*, vol. 6s1, p. BII.S11661, Jan. 2013, publisher: SAGE Publications Ltd STM. 22, 25

[130] J. Yuan, C. Holtz, T. Smith, and J. Luo, "Autism spectrum disorder detection from semi-structured and unstructured medical data," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2017, no. 1, p. 3, Feb. 2017. 22

[131] R. C. Bagot, B. Labonté, C. J. Peña, and E. J. Nestler, "Epigenetic signaling in psychiatric disorders: stress and depression," *Dialogues Clin Neurosci*, vol. 16, no. 3, pp. 281–295, Sep. 2014. 22

[132] Y.-Y. Liu, X.-Y. Zhou, L.-N. Yang, H.-Y. Wang, Y.-Q. Zhang, J.-C. Pu, L.-X. Liu, S.-W. Gui, L. Zeng, J.-J. Chen, C.-J. Zhou, and P. Xie, "Social defeat stress causes depression-like behavior with metabolite changes in the prefrontal cortex of rats," *PLOS ONE*, vol. 12, no. 4, p. e0176725, Apr. 2017, publisher: Public Library of Science. 22

[133] D. Smirnova, P. Cumming, E. Sloeva, N. Kuvshinova, D. Romanov, and G. Nosachev, "Language Patterns Discriminate Mild Depression From Normal Sadness and Euthymic State," *Front. Psychiatry*, vol. 9, 2018, publisher: Frontiers. 22, 24, 73

[134] J. S. Beck, *Cognitive therapy: Basics and beyond*. Guilford Press, New York, 1995. 22

[135] 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, Dec. 2004. 22, 39

[136] J. D. Bernard, J. L. Baddeley, B. F. Rodriguez, and P. A. Burke, "Depression, Language, and Affect: An Examination of the Influence of Baseline Depression and Affect Induction on Language," *Journal of Language and Social Psychology*, Jun. 2015, publisher: SAGE PublicationsSage CA: Los Angeles, CA. 22

[137] A. R. Sonnenschein, S. G. Hofmann, T. Ziegelmayer, and W. Lutz, "Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy," *Cognitive Behaviour Therapy*, vol. 47, no. 4, pp. 315–327, Jul. 2018. 23, 39

[138] M. D. Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media," *ICWSM*, vol. 7, no. 1, Jun. 2013, number: 1. 23

[139] J. Wolohan, M. Hiraga, A. Mukherjee, Z. A. Sayyed, and M. Millard, "Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with NLP," in *Proceedings of the First International Workshop on Language Cognition and Computational Models*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 11–21. 23

[140] A. C. Arevian, D. Bone, N. Malandrakis, V. R. Martinez, K. B. Wells, D. J. Miklowitz, and S. Narayanan, "Clinical state tracking in serious mental illness through computational analysis of speech," *PLoS ONE*, vol. 15, no. 1, p. e0225695, Jan. 2020. 24

[141] S. A. Qureshi, G. Dias, M. Hasanuzzaman, and S. Saha, "Improving Depression Level Estimation by Concurrently Learning Emotion Intensity," *IEEE Comput. Intell. Mag.*, vol. 15, no. 3, pp. 47–59, Aug. 2020. 24, 39, 135

[142] V. Rapcan, S. D'Arcy, S. Yeap, N. Afzal, J. Thakore, and R. B. Reilly, "Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia," *Medical Engineering & Physics*, vol. 32, no. 9, pp. 1074–1079, Nov. 2010. 25, 39, 134

[143] Y. Tahir, D. Chakraborty, J. Dauwels, N. Thalmann, D. Thalmann, and J. Lee, "Non-verbal speech analysis of interviews with schizophrenic patients," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 5810–5814. 25, 39, 134

[144] B. N. Axelrod, R. S. Goldman, and L. D. Alphs, "Validation of the 16-item negative symptom assessment," *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 253–258, Jul. 1993. 25

[145] A. S. Cohen, M. Alpert, T. M. Nienow, T. J. Dinzeo, and N. M. Docherty, "Computerized Measurement of Negative Symptoms in Schizophrenia," *J Psychiatr Res*, vol. 42, no. 10, pp. 827–836, Aug. 2008. 25

[146] A. S. Cohen, K. R. Mitchell, N. M. Docherty, and W. P. Horan, "Vocal expression in schizophrenia: Less than meets the ear," *J Abnorm Psychol*, vol. 125, no. 2, pp. 299–309, Feb. 2016. 26

[147] 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)*, Sep. 2017, pp. 000 213–000 218. 27

[148] T. F. Quatieri and N. Malyska, "Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity," *13th Annual Conference of the International Speech Communication Association*, pp. 1059–1062, Sep. 2012. 27, 39

[149] F. Honig, A. Batliner, E. Noth, S. Schnieder, and J. Krajewski, "Automatic Modelling of Depressed Speech: Relevant Features and Relevance of Gender," *15th Annual Conference of the International Speech Communication Association*, pp. 1248–1252, Sep. 2014. 27, 39

[150] 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, Jan. 2007. 27

[151] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt, "On the relative importance of vocal source, system, and prosody in human depression," in *2013 IEEE International Conference on Body Sensor Networks*, May 2013, pp. 1–6, iSSN: 2376-8894. 27

[152] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, Jul. 2000. 27

[153] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response," *Biological Psychiatry*, vol. 72, no. 7, pp. 580–587, Oct. 2012. 27

[154] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization," *INTERSPEECH 2011*, pp. 2997–3000, Aug. 2011. 27, 29

[155] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling Spectral Variability for the Classification of Depressed Speech," in *14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013): Speech in Life Sciences and Human Societies*. International Speech Communication Association, 2013, pp. 857–861. 29, 74

[156] 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)*, May 2019, pp. 5856–5860.

[157] Z. Huang, J. Epps, D. Joachim, B. Stasak, J. R. Williamson, and T. F. Quatieri, "Domain Adaptation for Enhancing Speech-Based Depression Detection in Natural Environmental Conditions Using Dilated CNNs," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 4561–4565. 27, 29, 39, 135

[158] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The Distress Analysis Interview Corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European

Language Resources Association (ELRA), May 2014, pp. 3123–3128. 27, 28, 95, 96

[159] S. Alghowinem, "From Joyous to Clinically Depressed: Mood Detection Using Multimodal Analysis of a Person's Appearance and Speech," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 648–654, iSSN: 2156-8111. 27, 29

[160] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting Depression Severity from Vocal Prosody," *IEEE Trans Affect Comput*, vol. 4, no. 2, pp. 142–150, 2013. 27, 30, 73

[161] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge," *arXiv:1605.01600 [cs]*, Nov. 2016, arXiv: 1605.01600. 27, 33, 39, 137

[162] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 3–12. 27, 97, 98

[163] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear, "Cross-cultural detection of depression from nonverbal behaviour," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, May 2015, pp. 1–8. 27, 39, 120, 136

[164] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. Geneva, Switzerland: IEEE, Sep. 2013, pp. 147–152. 28

[165] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "A Random Forest Regression Method With Selected-Text Feature For Depression Assessment," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC '17*. Mountain View, California, USA: ACM Press, 2017, pp. 61–68. 29

[166] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision Tree Based Depression Classification from Audio Video and Language Information," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 89–96. 29, 39, 136

[167] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber, M. E. Thase, J. H. Kocsis, and M. B. Keller, "The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression," *Biological Psychiatry*, vol. 54, no. 5, pp. 573–583, Sep. 2003. 29

[168] 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*. Vancouver, BC, Canada: IEEE, May 2013, pp. 7547–7551. 29

[169] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *J Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, Nov. 2013. 29, 39, 137

[170] B. Stasak, J. Epps, and R. Goecke, "Automatic depression classification based on affective read sentences: Opportunities for text-dependent analysis," *Speech Communication*, vol. 115, pp. 1–14, Dec. 2019. 30

[171] I. W. Miller, S. Bishop, W. H. Norman, and H. Maddever, "The modified Hamilton rating scale for depression: Reliability and validity," *Psychiatry Research*, vol. 14, no. 2, pp. 131–142, Feb. 1985. 30

[172] L. Albuquerque, A. R. S. Valente, A. Teixeira, D. Figueiredo, P. Sa-Couto, and C. Oliveira, "Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan," *PLoS One*, vol. 16, no. 4, Apr. 2021. 30

[173] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. dos Santos, "Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study," *Res. Biomed. Eng.*, vol. 37, no. 1, pp. 53–64, Mar. 2021. 30, 39, 135

[174] M. K. Mandal, R. Pandey, and A. B. Prasad, "Facial Expressions of Emotions and Schizophrenia: A Review," *Schizophr Bull*, vol. 24, no. 3, pp. 399–412, Jan. 1998. 30

[175] E. F. Walker, K. E. Grimes, D. M. Davis, and A. J. Smith, "Childhood precursors of schizophrenia: facial expressions of emotion." *The American journal of psychiatry*, 1993, publisher: American Psychiatric Assn. 31

[176] M.-y. CHU, X. LI, Q.-y. LV, Z.-h. Yl, E. F. C. CHEUNG, and R. C. K. CHAN, "Pleasure Experience and Emotion Expression in Patients with Schizophrenia," *Shanghai Arch Psychiatry*, vol. 29, no. 5, pp. 268–276, 2017. 31, 39

[177] C. C. Martin, J. C. Borod, M. Alpert, A. Brozgold, and J. Welkowitz, "Spontaneous expression of facial emotion in schizophrenic and right-brain-damaged patients," *Journal of Communication Disorders*, vol. 23, no. 4, pp. 287–301, Aug. 1990. 31

[178] F. Trémeau, "A review of emotion deficits in schizophrenia," *Dialogues Clin Neurosci*, vol. 8, no. 1, pp. 59–70, Mar. 2006. 31

[179] J. S. Lee, J. W. Chun, S. Y. Yoon, H.-J. Park, and J.-J. Kim, "Involvement of the mirror neuron system in blunted affect in schizophrenia," *Schizophrenia Research*, vol. 152, no. 1, pp. 268–274, Jan. 2014. 31

[180] J. Price, V. Cole, and G. M. Goodwin, "Emotional side-effects of selective serotonin reuptake inhibitors: qualitative study," *The British Journal of Psychiatry*, vol. 195, no. 3, pp. 211–217, Sep. 2009, publisher: Cambridge University Press. 31

[181] W. V. Friesen, P. Ekman, and others, "EMFACS-7: Emotional facial action coding system," *Unpublished manuscript, University of California at San Francisco*, vol. 2, no. 36, p. 1, 1983. 31

[182] P. Ekman, E. Rosenberg, and J. Hager, *Facial action coding system affect interpretation dictionary (FACSAID)*, 1998. 31

[183] C. Alvino, C. Kohler, F. Barrett, R. E. Gur, R. C. Gur, and R. Verma, "Computerized measurement of facial expression of emotions in schizophrenia," *Journal of Neuroscience Methods*, vol. 163, no. 2, pp. 350–361, Jul. 2007. 31

[184] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall, "Automated Facial Expressions Analysis in Schizophrenia: A Continuous Dynamic Approach," in *Pervasive Computing Paradigms for Mental Health*. Springer, Cham, Sep. 2015, pp. 72–81. 32, 39, 134, 139

[185] M. Bishay, S. Priebe, and I. Patras, "Can Automatic Facial Expression Analysis Be Used for Treatment Outcome Estimation in Schizophrenia?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, May 2019, pp. 1632–1636. 32, 138

[186] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. Amsterdam, Netherlands: IEEE, Sep. 2009, pp. 1–7. 32, 39, 136

[187] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, Oct. 2014. 32

[188] J. M. Girard and J. F. Cohn, "Automated audiovisual depression analysis," *Current Opinion in Psychology*, vol. 4, pp. 75–79, Aug. 2015. 32, 33, 39

[189] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, Mar. 2018, conference Name: IEEE Journal of Biomedical and Health Informatics. 33, 39, 120, 136

[190] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear, "Head Pose and Movement Analysis as an Indicator of Depression," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 283–288, iSSN: 2156-8111. 33, 39

[191] Y. Wang, J. Ma, B. Hao, P. Hu, X. Wang, J. Mei, and S. Li, "Automatic Depression Detection via Facial Expressions Using Multiple Instance Learning," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2020, pp. 1933–1936, iSSN: 1945-8452. 33

[192] D. Aspandi, A. Mallol-Ragolta, B. Schuller, and X. Binefa, "Latent-Based Adversarial Neural Networks for Facial Affect Estimations," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, Nov. 2020, pp. 606–610.

[193] X. Li, W. Guo, and H. Yang, "Depression severity prediction from facial expression based on the DRR_depressionnet network," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2020, pp. 2757–2764.

[194] W. Carneiro de Melo, E. Granger, and M. Bordallo Lopez, "MDN: A Deep Maximization-Differentiation Network for Spatio-Temporal Depression Detection," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021, conference Name: IEEE Transactions on Affective Computing.

[195] W. Guo, H. Yang, Z. Liu, Y. Xu, and B. Hu, "Deep Neural Networks for Depression Recognition Based on 2D and 3D Facial Expressions Under Emotional Stimulus Tasks," *Front. Neurosci.*, vol. 15, 2021, publisher: Frontiers. 33

[196] M. Morrens, L. Docx, and S. Walther, "Beyond Boundaries: In Search of an Integrative View on Motor Symptoms in Schizophrenia," *Front. Psychiatry*, vol. 5, 2014, publisher: Frontiers. 33

[197] F. Ramseyer and W. Tschacher, "Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome." *Journal of Consulting and Clinical Psychology*, vol. 79, no. 3, pp. 284–295, 2011. 34

[198] F. Ramseyer, "Motion energy analysis (MEA): A primer on the assessment of motion from video," *Journal of counseling psychology*, vol. 67, pp. 536–549, Jul. 2020. 34, 39

[199] L. L. Dutschke, K. Stegmayer, F. Ramseyer, S. Bohlhalter, T. Vanbellingen, W. Strik, and S. Walther, "Gesture impairments in schizophrenia are linked to increased movement and prolonged motor planning and execution," *Schizophrenia Research*, vol. 200, pp. 42–49, Oct. 2018. 34, 39

[200] M. Lavelle, P. G. T. Healey, and R. McCabe, "Is Nonverbal Communication Disrupted in Interactions Involving Patients With Schizophrenia?" *Schizophr Bull*, vol. 39, no. 5, pp. 1150–1158, Sep. 2013. 34

[201] A. Vaskinn, K. Sundet, T. Østefjells, K. Nymo, I. Melle, and T. Ueland, "Reading Emotions from Body Movement: A Generalized Impairment in Schizophrenia," *Front. Psychol.*, vol. 6, 2016. 34

[202] U. Altmann, M. Brümmel, J. Meier, and B. Strauss, "Movement Synchrony and Facial Synchrony as Diagnostic Features of Depression: A Pilot Study," *The Journal of Nervous and Mental Disease*, vol. 209, no. 2, pp. 128–136, Feb. 2021. 34, 39

[203] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, Oct. 2010. 35

[204] A. Weise and R. Levitan, "Looking for Structure in Lexical and Acoustic-Prosodic Entrainment Behaviors," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 297–302. 35

[205] S. Kousidis, D. Dorran, Y. Wang, B. Vaughan, C. Cullen, D. Campbell, C. McDonnell, and E. Coyle, "Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues," Sep. 2008. 36

[206] B. Vaughan, "Prosodic Synchrony in Co-operative Task-based Dialogues: A Measure of Agreement and Disagreement," pp. 1865–1868, Aug. 2011. 36

[207] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, pp. 11–34, Mar. 2014. 36

[208] B. Vaughan, C. De Pasquale, L. Wilson, C. Cullen, and B. Lawlor, "Investigating Prosodic Accommodation in Clinical Interviews with Depressed Patients," in *Pervasive Computing Paradigms for Mental Health*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, P. Cipresso, S. Serino, Y. Ostrovsky, and J. T. Baker, Eds. Springer International Publishing, 2018, pp. 150–159. 36, 37

[209] C. D. Pasquale, C. Cullen, and B. Vaughan, "An Investigation of Therapeutic Rapport Through Prosody in Brief Psychodynamic Psychotherapy," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 3043–3047. 36

[210] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, Mar. 2014. 36

[211] M. M. Willi, S. A. Borrie, T. S. Barrett, M. Tu, and V. Berisha, "A Discriminative Acoustic-Prosodic Approach for Measuring Local Entrainment," *arXiv:1804.08663 [cs, eess]*, Jul. 2018, arXiv: 1804.08663. 36

[212] B. Xiao, P. Georgiou, Z. E. Imel, D. C. Atkins, and S. S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addiction counseling," in *INTERSPEECH*, Aug. 2013, pp. 2861–2865. 37

[213] M. Nasir, B. Baucom, C. J. Bryan, S. S. Narayanan, and P. Georgiou, "Complexity in Speech and its Relation to Emotional Bond in Therapist-Patient Interactions During Suicide Risk Assessment Interviews," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 3296–3300. 37

[214] B. Elvevåg, P. W. Foltz, D. R. Weinberger, and T. E. Goldberg, "Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia," *Schizophr Res*, vol. 93, no. 1-3, pp. 304–316, Jul. 2007. 39, 133

[215] B. Elvevåg, P. W. Foltz, M. Rosenstein, and L. E. DeLisi, "An automated method to analyze language use in patients with schizophrenia and their first-degree relatives," *Journal of Neurolinguistics*, vol. 23, no. 3, pp. 270–284, May 2010. 39, 133

[216] A. Voppel, J. de Boer, S. Brederoo, H. Schnack, and I. Sommer, "Quantified language connectedness in schizophrenia-spectrum disorders," *Psychiatry Research*, vol. 304, p. 114130, Oct. 2021, 74. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165178121004261 39

[217] N. Rezaii, E. Walker, and P. Wolff, "A machine learning approach to predicting psychosis using semantic density and latent content analysis," *npj Schizophr*, vol. 5, no. 1, p. 9, Dec. 2019. 39, 74, 119, 133

[218] R. Martínez-Castaño, J. C. Pichel, and D. E. Losada, "A Big Data Platform for Real Time Analysis of Signs of Depression in Social Media," *IJERPH*, vol. 17, no. 13, p. 4752, Jul. 2020. 39

[219] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. d. Santos, "Vocal acoustic analysis and machine learning for the identification of schizophrenia," *Res. Biomed. Eng.*, vol. 37, no. 1, pp. 33–46, Mar. 2021, company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Springer International Publishing. 39

[220] Y. Lu, A. Harati, T. Rutowski, R. Oliveira, P. Chlebek, and E. Shriberg, "Robust Speech and Natural Language Processing Models for Depression Screening," in *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Dec. 2020, pp. 1–5, 60. 39

[221] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD," *INTERSPEECH-2013*, pp. 847–851, Aug. 2013. 39

[222] A. Harati, E. Shriberg, T. Rutowski, P. Chlebek, Y. Lu, and R. Oliveira, "Speech-Based Depression Prediction Using Encoder-Weight-Only Transfer Learning and a Large Corpus," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 7273–7277, 55. 39

[223] K. E. B. Ooi, M. Lech, and N. B. Allen, "Multichannel Weighted Speech Classification System for Prediction of Major Depression in Adolescents," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 2, pp. 497–506, Feb. 2013. 39, 135

[224] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. F. Cohn, "Multimodal Detection of Depression in Clinical Interviews," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 307–310, event-place: Seattle, Washington, USA. 39, 136

[225] L.-S. A. Low, M. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of Clinical Depression in Adolescents' Speech During Family Interactions," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, Mar. 2011. 39, 120, 135

[226] M. H. Sanchez, D. Vergyri, L. Ferrer, C. Richey, P. Garcia, B. Knoth, and W. Jarrold, "Using prosodic and spectral features in detecting depression in elderly males," in *12th Annual Conference of the International Speech Communication Association*, Aug. 2011, pp. 3001–3004. 39, 135

[227] T. Taguchi, "Major depressive disorder discrimination using vocal acoustic features," *Journal of Affective Disorders*, p. 7, 2018. 39, 135

[228] B. Buck, K. S. Minor, and P. H. Lysaker, "Lexical Characteristics of Anticipatory and Consummatory Anhedonia in Schizophrenia: A Study of Language in Spontaneous Life Narratives," *Journal of Clinical Psychology*, vol. 71, no. 7, pp. 696–706, 2015. 39

[229] K. A. Bonfils, L. Luther, R. L. Firmin, P. H. Lysaker, K. S. Minor, and M. P. Salyers, "Language and hope in schizophrenia-spectrum disorders," *Psychiatry Research*, vol. 245, pp. 8–14, Nov. 2016.

[230] N. B. Mota, N. A. P. Vasconcelos, N. Lemos, A. C. Pieretti, O. Kinouchi, G. A. Cecchi, M. Copelli, and S. Ribeiro, "Speech Graphs Provide a Quantitative Measure of Thought Disorder in Psychosis," *PLoS ONE*, vol. 7, no. 4, p. e34928, Apr. 2012. 39, 119, 133

[231] R. Kliper, S. Portuguese, and D. Weinshall, "Prosodic Analysis of Speech and the Underlying Mental State," in *Pervasive Computing Paradigms for Mental Health*. Springer, Cham, 2016, pp. 52–62. [Online]. Available: https://link.springer.com.remotexs.ntu.edu.sg/chapter/10.1007/978-3-319-32270-4_6 39

[232] C. Perlini, A. Marini, M. Garzitto, M. Isola, S. Cerruti, V. Marinelli, G. Rambaldelli, A. Ferro, L. Tomelleri, N. Dusi, M. Bellani, M. Tansella, F. Fabbro, and P. Brambilla, "Linguistic production and syntactic comprehension in schizophrenia and bipolar disorder," *Acta Psychiatrica Scandinavica*, vol. 126, no. 5, pp. 363–376, 2012.

[233] G. Kiss and K. Vicsi, "Mono- and multi-lingual depression prediction based on speech processing," *Int J Speech Technol*, vol. 20, no. 4, Dec. 2017. [Online]. Available: https://link.springer.com.remotexs.ntu.edu.sg/article/10.1007/s10772-017-9455-8 39

[234] L. M. Bylsma, B. H. Morris, and J. Rottenberg, "A meta-analysis of emotional reactivity in major depressive disorder," *Clinical Psychology Review*, vol. 28, no. 4, pp. 676–691, Apr. 2008. 39, 94

[235] O. Puig, R. Penadés, I. Baeza, E. De la Serna, V. Sánchez-Gistau, M. Bernardo, and J. Castro-Fornieles, "Cognitive Remediation Therapy in Adolescents With Early-Onset Schizophrenia: A Randomized Controlled Trial," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 53, no. 8, pp. 859–868, Aug. 2014. 42

[236] M. B. First, R. L. Spitzer, M. Gibbon, J. B. Williams, and others, "Structured clinical interview for DSM-IV-TR axis I disorders, research version, patient edition," SCID-I/P New York, NY, USA:, Tech. Rep., 2002. 43

[237] R. Keefe, "The Brief Assessment of Cognition in Schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery," *Schizophrenia Research*, vol. 68, no. 2-3, pp. 283–297, Jun. 2004. 45, 122

[238] G. Rekhi, L. Alphs, M. S. Ang, and J. Lee, "Clinical utility of the Negative Symptom Assessment-16 in individuals with schizophrenia," *European Neuropsychopharmacology*, vol. 29, no. 12, pp. 1433–1441, Dec. 2019. 46

[239] A. Shafer, "Meta-Analysis of the Brief Psychiatric Rating Scale Factor Structure," *Psychological assessment*, vol. 17, no. 3, pp. 324–335, 2005, place: Washington, DC Publisher: American Psychological Association, American Psychological Association, Inc. 46

[240] S. Leucht, J. Kane, W. Kissling, J. Hamann, E. Etschel, and R. Engel, "What does the PANSS mean?" *Schizophrenia Research*, vol. 79, no. 2-3, pp. 231–238, Nov. 2005. 50, 91

[241] S. Leucht, J. M. Kane, E. Etschel, W. Kissling, J. Hamann, and R. R. Engel, "Linking the PANSS, BPRS, and CGI: Clinical Implications," *Neuropsychopharmacology*, vol. 31, no. 10, pp. 2318–2325, Oct. 2006, number: 10 Publisher: Nature Publishing Group.

[242] S. Leucht, Barabássy, I. Laszlovszky, B. Szatmári, K. Acsai, E. Szalai, J. Harsányi, W. Earley, and G. Németh, "Linking PANSS negative symptom scores with the Clinical Global Impressions Scale: understanding negative symptom scores in schizophrenia," *Neuropsychopharmacology*, vol. 44, no. 9, pp. 1589–1596, Aug. 2019. 50, 91

[243] Z. Yang, K. Lim, M. Lam, R. Keefe, and J. Lee, "Factor structure of the positive and negative syndrome scale (PANSS) in people at ultra high risk (UHR) for psychosis," *Schizophrenia Research*, vol. 201, pp. 85–90, Nov. 2018. 50

[244] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and others, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011, publisher: JMLR. org. 50, 53, 55

[245] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002. 52

[246] 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, p. e0118432, Mar. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/ 54

[247] C. Molnar, *Interpretable Machine Learning*. Lulu.com, Feb. 2020, google-Books-ID: jBm3DwAAQBAJ. 55

[248] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, Oct. 2010, pp. 1459–1462. 56, 60

[249] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei, A. Vann, N. Vogler, T. Bocklet, M. Cernak, J. Hannink, and E. Nöth, "NeuroSpeech: An open-source software for Parkinson's speech analysis," *Digital Signal Processing*, vol. 77, pp. 207–221, Jun. 2018. 56

[250] P. Soille, *Morphological Image Analysis: Principles and Applications*. Springer Science & Business Media, Mar. 2013, google-Books-ID: ZFzxCAAAQBAJ. 57

[251] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "JHU ASpIRE system: Robust LVCSR with TDNNS, iVector adaptation and RNN-LMS," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Scottsdale, AZ, USA: IEEE, Dec. 2015, pp. 539–546. 58

[252] R. Řeh\uuřek, P. Sojka, and others, "Gensim—statistical semantics in python," *Retrieved from genism. org*, 2011. 59

[253] J. C. Vásquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. Nöth, "Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech," in *INTERSPEECH*, 2019. 60, 61

[254] W. Jarrold, H. S. Javitz, R. Krasnow, B. Peintner, E. Yeh, G. E. Swan, and M. Mehl, "Depression and Self-Focused Language in Structured Interviews with Older Men," *Psychol Rep*, vol. 109, no. 2, pp. 686–700, Oct. 2011. 73

[255] C. Lambert, S. D. Silva, A. K. Ceniti, S. J. Rizvi, G. Foussias, and S. H. Kennedy, "Anhedonia in depression and schizophrenia: A transdiagnostic challenge," *CNS Neuroscience & Therapeutics*, vol. 24, no. 7, pp. 615–623, 2018, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cns.12854. 73

[256] H. Jiang, B. Hu, Z. Liu, G. Wang, L. Zhang, X. Li, and H. Kang, "Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech Features," *Comput Math Methods Med*, vol. 2018, Sep. 2018. 74

[257] 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, Dec. 2015. 74

[258] M. L. Birnbaum, S. K. Ernala, A. F. Rizvi, E. Arenare, A. R. Van Meter, M. De Choudhury, and J. M. Kane, "Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook," *npj Schizophrenia*, vol. 5, no. 1, pp. 1–9, Oct. 2019, number: 1 Publisher: Nature Publishing Group. 74

[259] J. Zhang, Z. Pan, C. Gui, J. Zhu, and D. Cui, "Clinical investigation of speech signal features among patients with schizophrenia," *Shanghai Arch Psychiatry*, vol. 28, no. 2, pp. 95–102, Apr. 2016. 75, 121

[260] D. McDuff, R. el Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected "In-the-Wild"," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2013, pp. 881–888, iSSN: 2160-7516. 78

[261] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 59–66. 78

[262] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015, arXiv: 1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556 97

[263] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *arXiv:1608.06993 [cs]*, Jan. 2018, arXiv: 1608.06993. [Online]. Available: http://arxiv.org/abs/1608.06993 97

[264] J. Fu, S. Yang, F. He, L. He, Y. Li, J. Zhang, and X. Xiong, "Sch-net: a deep learning architecture for automatic detection of schizophrenia," *BioMedical Engineering OnLine*, vol. 20, no. 1, p. 75, Aug. 2021. [Online]. Available: https://doi.org/10.1186/s12938-021-00915-2 97, 98

[265] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond," *Int J Comput Vis*, vol. 127, no. 6, pp. 907–929, Jun. 2019. [Online]. Available: https://doi.org/10.1007/s11263-019-01158-4 97

[266] R. EBU–Recommendation, "Loudness normalisation and permitted maximum level of audio signals," 2011, publisher: Citeseer. 102

[267] R. Levitan and J. Hirschberg, "Measuring Acoustic-Prosodic Entrainment with Respect to Multiple Levels and Dimensions," p. 4, 2011. 103

[268] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia - MM '13*. Barcelona, Spain: ACM Press, 2013, pp. 835–838. 103

[269] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016. 104

[270] N. J. Enfield, *How We Talk: the inner workings of conversation*, first edition ed. New York: Basic Books, 2017, oCLC: 975485829. 107

[271] T. Wörtwein, T. Baltrušaitis, E. Laksana, L. Pennant, E. S. Liebson, D. Öngür, J. T. Baker, and L.-P. Morency, "Computational Analysis of Acoustic Descriptors in Psychotic Patients," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 3256–3260. 122, 138, 139

[272] S. Kumar, A. Veldhuis, and T. Malhotra, "Neuropsychiatric and Cognitive Sequelae of COVID-19," *Front. Psychol.*, vol. 12, 2021, publisher: Frontiers. 125

[273] G. P. Strauss, K. I. Macdonald, I. Ruiz, I. M. Raugh, L. A. Bartolomeo, and S. H. James, "The impact of the COVID-19 pandemic on negative symptoms in individuals at clinical high-risk for psychosis and outpatients with chronic schizophrenia," *Eur Arch Psychiatry Clin Neurosci*, Apr. 2021. 125

[274] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Conference on Fairness, Accountability and Transparency*. PMLR, Jan. 2018, pp. 77–91, iSSN: 2640-3498. 126

[275] E. Granholm, J. L. Holden, T. Mikhael, P. C. Link, J. Swendsen, C. Depp, R. C. Moore, and P. D. Harvey, "What Do People With Schizophrenia Do All Day? Ecological Momentary Assessment of Real-World Functioning in Schizophrenia," *Schizophrenia Bulletin*, vol. 46, no. 2, pp. 242–251, Feb. 2020. [Online]. Available: https://doi.org/10.1093/schbul/sbz070 127

[276] C. A. Depp, J. Bashem, R. C. Moore, J. L. Holden, T. Mikhael, J. Swendsen, P. D. Harvey, and E. L. Granholm, "GPS mobility as a digital biomarker of negative symptoms in schizophrenia: a case control study," *npj Digit. Med.*, vol. 2, no. 1, pp. 1–7, Nov. 2019, bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Health care;Medical research Subject_term_id: health-care;medical-research. [Online]. Available: https://www.nature.com/articles/s41746-019-0182-1 127

[277] H.-y. Zhou, X.-l. Cui, B.-r. Yang, L.-j. Shi, X.-r. Luo, E. F. C. Cheung, S. S. Y. Lui, and R. C. K. Chan, "Audiovisual Temporal Processing in Children and Adolescents With Schizophrenia and Children and Adolescents With Autism: Evidence From Simultaneity-Judgment Tasks and Eye-Tracking Data:," *Clinical Psychological Science*, Jul. 2021, publisher: SAGE PublicationsSage CA: Los Angeles, CA. [Online]. Available: https://journals.sagepub.com.remotexs.ntu.edu.sg/doi/full/10.1177/21677026211031543 128

[278] H.-y. Zhou, X.-l. Cai, M. Weigl, P. Bang, E. F. C. Cheung, and R. C. K. Chan, "Multisensory temporal binding window in autism spectrum disorders and schizophrenia spectrum disorders: A systematic review and meta-analysis," *Neuroscience & Biobehavioral Reviews*, vol. 86, pp. 66–76, Mar. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763417307625 128

[279] R. A. Stevenson, S. Park, C. Cochran, L. G. McIntosh, J.-P. Noel, M. D. Barense, S. Ferber, and M. T. Wallace, "The associations between multisensory temporal processing and symptoms of schizophrenia," *Schizophrenia Research*, vol. 179, pp. 97–103, Jan. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920996416304443 128

[280] R. A. Stevenson, M. Segers, B. L. Ncube, K. R. Black, J. M. Bebko, S. Ferber, and M. D. Barense, "The cascading influence of multisensory processing on speech perception in autism," *Autism*, vol. 22, no. 5, pp. 609–624, 2018, publisher: Sage Publications Sage UK: London, England. 128

[281] S. Xu, Z. Yang, D. Chakraborty, Y. Tahir, T. Maszczyk, Y. H. V. Chua, J. Dauwels, D. Thalmann, N. M. Thalmann, B.-L. Tan, and J. L. C. Keong, "Automated Lexical Analysis of Interviews with Individuals with Schizophrenia," in *9th International Workshop on Spoken Dialogue System Technology*, L. F. D'Haro, R. E. Banchs, and H. Li, Eds. Singapore: Springer Singapore, 2019, pp. 185–197. 133

[282] F. Martínez-Sánchez, J. A. Muela-Martínez, P. Cortés-Soto, J. J. García Meilán, J. A. Vera Ferrándiz, A. Egea Caparrós, and I. M. Pujante Valverde, "Can the Acoustic Analysis of Expressive Prosody Discriminate Schizophrenia?" *Span. J. Psychol.*, vol. 18, p. E86, 2015. 134

[283] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Apr. 2013, pp. 1–8. 135