Deep Learning-based Multi-modal Depression Estimation using Knowledge from Micro-facial Expressions, Audio and Text

Bachelor's thesis

by

Ping-Cheng Wei, B.Sc.

at

Institute of Industrial Information Technology

Time period:	22/11/2021 – 28/04/2022
Supervisor:	Prof. Dr.-Ing. Michael Heizmann[1]
	Prof. Dr.-Ing. Rainer Stiefelhagen[2]
Advisor:	M.Sc. Kunyu Peng[2]
	Dr.-Ing. Alina Roitberg[2]

[1]Institute of Industrial Information Technology (IIIT)
[2]Institute for Anthropomatics and Robotics (IAR)

Declaration

I hereby declare that I wrote my Bachelor's thesis on my own and that I have followed the regulations relating to good scientific practice of the Karlsruhe Institute of Technology (KIT) in its latest form. I did not use any unacknowledged sources or means and I marked all references I used literally or by content.

Karlsruhe, 28 April 2022

Kurzfassung

Depression ist eine psychiatrische Störung, die das Leben und die Familie eines persons negativ beeinflussen kann. Aufgrund der Barrieren und Einschränkungen für eine rechtzeitige Diagnose und wirksame Behandlung leiden weltweit mehr als 280 Millionen Menschen an Depressionen. Dennoch erhalten nur weniger als 40% der Erwachsener mit psychischen Erkrankungen eine angemessene Behandlung. Wenn die Depression schlimmer wird, kann sie dazu führen, dass eine Person Selbstmord begeht. In Deutschland etwa sind etwa 60% aller Suizide auf Depressionen zurückzuführen. Um die Grundursache eines solchen Problems zu lösen, zielte diese Arbeit daher darauf ab, ein automatisches Deep Learning-basiertes Modell zur Schätzung von Depressionen zu realisieren. Genauer gesagt wird ein multimodales Modell erwartet, da man sich vorstellen kann, dass die Depressionssymptome nicht nur in einer Modalität, sondern in Multimodalität ausgedrückt werden.

Für die Multimodalität werden in dieser Arbeit visuelle Daten des Mikrogesichtsausdrucks, Audiodaten des Log-Mel Spektrogramms und Textdaten der Satzeinbettung verwendet. Da die Rohdaten im DAIC-WOZ jedoch verschiedene Fehler und potenzielle Probleme enthalten, wurden einige Feature-Engineering-Techniken implementiert, wie z. B. die Sliding-Window-Technik und das Gender-Balancing. Der endgültig generierte geclippte Datensatz liefert einen großen Erfolg in allen Experimenten, indem er eine Leistungsverbesserung von mindestens 20% zeigt, was auf die Bedeutung der Sauberkeit eines Datensatzes hinweist.

Zur Abschätzung der Schwere von Depressionen sowie des binären Zustands einer schweren Depression stellt diese Arbeit eine neuartige Idee zur Anwendung eines Subaufmerksamkeitsmechanismus zusammen mit einer teilnehmerbasierten Analyse vor, die zu einer endgültigen Vorhersagegenauigkeit von 85,11% für Major Depression und eine Genauigkeit von etwa 90% führt. Insgesamt sind unsere Ergebnisse mit den modernsten Methoden vergleichbar, und diese Arbeit zeigt eine erfolgreiche Kombination aus natürlicher Sprachverarbeitung, Computervision und Sprachverarbeitungstechniken, um tiefere zugrunde liegende Depressionshinweise für die Depressionsschätzung zu gewinnen.

Abstract

Depression is a psychiatric disorder that can negatively affect one's life and family. Owing to the barriers as well as limitations to timely diagnosis and effective care, worldwide, more than 280 million people are suffering from depression. Nevertheless, only less than 40% of audits with mental illness receive proper treatment. When depression gets more severe, it can lead a person to commit suicide. In Germany, for instance, about 60% of all suicides are due to depression. Therefore, to solve the root cause of such an issue, this thesis aimed at realizing an automatic deep learning-based model for depression estimation. More precisely, a multi-modal model is expected as one can imagine that the depression symptoms will not only be expressed in solely one modality but in multimodality.

For multimodality, visual data of micro-facial expression, audio data of log-mel spectrogram, and text data of sentence embedding are exploited in this work. However, since the raw data in the DAIC-WOZ involves various errors and potential problems, some feature engineering techniques have been implemented such as the sliding window technique and gender balancing. The final generated clipped dataset yields a great success in all the experiments by demonstrating at least a 20% performance improvement, indicating the importance of the cleanness of a dataset.

For estimating the severity of depression as well as the binary state of major depression, this work presents a novel idea of applying sub-attentional mechanism, along with participant-based analysis, which leads to a final 85.11% prediction accuracy for major depression and an around 90% precision. Overall, our results are comparable with the state-of-the-art methods and this work demonstrates a successful combination of natural language processing, computer vision, and speech processing techniques for harvesting deeper underlying depression cues for depression estimation.

Contents

List of Figures

List of Tables

Abbreviations and Symbols

Abbreviations

Abbreviation	Meaning
MD	Major Depression
MDD	Major Depressive Disorder
PTSD	Post-Traumatic Stress Disorder
DAIC	Distress Analysis Interview Corpus
WOZ	Wizard-of-Oz
PHQ	Patient Health Questionnaire
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, fourth edition
GT	Ground Truth
ML	Machine Learning
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
F0	Fundamental Frequency
NAQ	Normalized Amplitude Quotient
GB	Gender Balancing
FC	Fully Connected Layer
SVM	Support Vector Machine
CNN	Convolutional Neural Network
STFT	Short-Time Fourier Transformation
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
MUSDL	Multi-path Uncertainty-aware Score Distributions Learning
CV	Computer Vision
ViT	Vision Transformer
SAM	Sharpness-Aware Minimization
NLP	Natural Language Processing
GCN	Graph Convolutional Networks
ST-GCN	Spatial-Temporal Graph Convolutional Networks

Latin letters

Symbol	Meaning
t	Zeit
$\mathbf{x}$	Signalvektor

Calligraphic and other symbols

Symbol	Meaning
$\mathbb{R}$	Menge der reellen Zahlen
$\|.\|_2$	Euklidische Norm
$\|.\|$	Matrixnorm
$\max$	Maximum
$\min$	Minimum
Im$\{\cdot\}$	Imaginärteil
Re$\{\cdot\}$	Realteil

Chapter 1

Introduction

Depression is a common and serious medical illness that negatively affects the way you think, how you feel, and how you behave. The severe stage of depression is called Major Depressive Disorder (MDD) or Major Depression (MD) in short, which is a psychiatric disorder defined as a mental state of pervasive and persistent low mood, accompanied by the possibility of aversion to activity. Neurophysiologically, MD results in changes in neurocognition that affect the control of linguistic, motor, and cognitive functions. In life, MD negatively impacts one's daily life and sense of well-being, inducing a person to become pessimistic, reluctant, and cynical. These lead to low self-esteem, loss of interest, and incapability, which indirectly ruin one's life, work, family, general health, eating habits, and sleep quality. In extreme conditions, a person might even commit suicide. Fortunately, depression is treatable if the person gets treated in time. Therefore, knowing the psychological condition of a person is urgent as people nowadays suffer even more from multiple pressure in their day-to-day life and an automatic depression estimation model can come in handy in this case, serving as an assistance device.

During Audio-Visual Emotion Challenge (AVEC) in 2017, deep learning approaches showed impressive performance for depression estimation. Several deep learning models have demonstrated promising results. Recently, various data-driven methods such as fusion of multimodality, extracting statistical data with intense feature engineering, and topic modeling have also been proposed and researched to further improve the performance of depression estimation. Therefore, training a deep learning-based multi-modal model comes to be a potential solution for depression estimation. In addition to that, micro-facial expression through facial key points and gaze direction might even help to harvest the deeper underlying depressive characteristics since depressive features are strongly related to the subtle facial expression changes.

1.1 Motivation

Without a doubt, depression is the most prevalent mental health disorder and the worldwide leading cause of disability. According to the World Health Organization (WHO), more than 280 million people in the world have depression [2]. In Germany itself, around 11% of the population suffers from depression [32, 22]. When depression develops into MD, it can lead a person to commit suicide. Based on the statistic from WHO, around 700,000 people decide to end their life due to the MD every year [2]. In Germany, on the other hand, more people died from suicide than from drugs, traffic accidents, and HIV combined [32]. 15% of patients with MD die from suicide and about 60% of all suicides are due to depression.

The underlying issue behind these terrifying statistics is that compared to physical illnesses, mental disorders such as depression are hard to detect or even got ignored. According to [22], around one-third (33%) of patients with depression are not recognized in the clinical diagnosis. Besides this, depression can easily be confused with other illnesses due to similar symptoms. In general, depression can be categorized into 8 major depression symptoms, i.e., loss of interest, depression, sleep problems, loss of energy, poor appetite, feeling restless, trouble concentrating, and feeling worthlessness. some of those symptoms also resemble the symptoms of melancholy. Furthermore, the burden of mental health is exacerbated by limitations and barriers to effective and timely care, including financial cost, lack of resources or well-trained experts, and social stigma.

Therefore, to address those problems, realizing automatic depression estimation through a multi-modal deep learning-based approach has been called for. If the result is successful and promising, it could significantly benefit tons of adults who are suffering from depression but do not receive the treatment.

1.2 Objective

The main focus of this work is to realize an automatic depression estimation model, particularly through a multi-modal deep learning-based approach. The possibility of leveraging multi-modal data to harvest deeper underlying depression cues in comparison to the single-modal model is also explored. Furthermore, the applicability of utilizing micro-facial expressions such as facial key points and gaze direction for depression detection is also wanted to be proven. Finally, the most effective and low-memory cost backbone as feature extractor and fusion method will be researched.

Chapter 2

Related Works and Background

Since multi-modal depression estimation is a very broad topic, including data preprocessing such as acoustic features and text features, deep learning model building, and effective backbone model discovery. Several books, papers, and literature have been researched. In this chapter, all the state-of-the-art related works to depression estimation are summarized in the Section 2.1.

2.1 Related Work

The current state-of-the-art research in terms of depression estimation can be categorized into the following 4 different groups: statistical ML approach, pure computer vision (CV) approach, natural language processing (NLP) approach, and a mixture of multiple approaches. Each approach has its advantage and disadvantage as well as simplicity and difficulties, which will be discussed in the following:

Statistical ML approach: What the statistical ML approach basically means is that instead of exploiting complex and high computational load models such as the deep learning model, the author focuses on implementing intensive feature engineering techniques to generate a cleaner depression dataset, which will then be trained with the basic statistical ML models, e.g. Support Vector Machine (SVM). Valstar et al. and Williamson et al. [37, 38], for example, implemented several feature engineering techniques to extract various visual and acoustic features, namely histogram of oriented gradients feature of the facial area, facial action units, prosodic features like Fundamental frequency (F0), and acoustic spectral features. These extracted features are then trained with the SVM model to make a binary classification of depression state and a regression for depression severity. Dham et al. [12] investigated the other methods of feature engineering by further deriving new statistics according to the raw depression data. For instance, with facial key points, the head motion, key points distance, blinking rate, etc. are being computed and for text data, the average words and number of sentences are calculated, which are then also trained with the SVM or essential neural network model.

Pure CV approach: For the pure CV approach, the author employs solely CNN-like model for depression estimation. Visual data input here is normally the RGB image or facial key points and audio data input here is the extracted spectrogram or mel spectrogram, which is considered a gray-scale signal image. With the pure CV approach, the author focuses more on model-driven methods rather than data-driven methods. Therefore, usually, the raw data is utilized. Haque et

al. [19] trained a Causal Convolutional Networks (C-CNN) with a multi-modal model based on a mixture of the raw audio spectrogram, visual, and text data and achieved a highly accurate performance. Saidi et al. [31] exploited CNN as a feature extractor combined with SVM as the output head for binary classification.

NLP approach: Since the depression database is lots of interview recordings of different participants, it can be considered as time-series data. Therefore, without a doubt, the NLP model will also come into question as it is specifically designed for processing sequence data. Ma et al. [30] created a model called "DepAudioNet", which is a combination of the 1D-CNN and an LSTM model. DepAudioNet was one of the best models in AVEC 2016 by exploiting a deep learning-based approach with solely acoustic features for depression detection. Currently, Zhao et al. [40] applied the multi-head attentional mechanism together with the LSTM model to predict the binary state of depression. Their model also achieved a new state-of-the-art performance.

mixture of multiple approaches: Finally, as for the mixture of multiple approaches, the author built the multi-modal model based on the different feature extractors which suit different input modalities the best, accordingly. With this method, the advantages of different backbones can be maximumly utilized and thus leads to the highest model performance. Lin et al. [26] employed CNN to extract the acoustic features from audio mel spectrogram and BiLSTM to extract text features, separately, and fused the extracted features to predict the severity of depression. Lam et al. [25], on the other hand, exploited the Transformer model for processing the text features, along with CNN for audio data.

In this work, our approach resembles the mixture of multiple approaches as well as the NLP approach. By combining the CNN layer with the BiLSTM layer together, a ConvBiLSTM model is created, which is then applied in each branch in a multi-modal model to extract the spatial and temporal features from each interview.

2.2 Sentence Embedding

Since computer can not process texts or sentences, a transformation is need to convert and represent them in a language that a computer could be able to process and understand, i.e. numbers or vectors. This technique of representing text as numbers is so-called text embedding. Initially, text embedding was only designed to convert a set of given words into a set of vectors, which is called word embedding. For instance, given a sentence of 6 words like "It is nice to meet you", it will be transformed into 6 vectors with each having a dimension of embedding dimension. However, with word embedding, the meaning of a sentence and the relationship as well as correlation between each word in that sentence will be deprived as it only convert the sentence in a word by word manner rather than the whole sentence. Besides that, what if, there are plenty of sentences rather just one sentence? In this case, converting only words with word embedding will be very tedious and limited by the information that could be retrieved since a

certain amount of context will lost in this manner. Therefore, instead of dealing with individual words, sentence embedding was then developed to work directly with each individual sentences retaining the meaning and correlation. Sentence embedding techniques represent entire sentences and their semantic information as vectors. This helps the computer in understanding the context, intention, and other nuances in the entire text.

Chapter 3

Depression Dataset

In this work, the DAIC-WOZ depression database [9] compiled by the USC's Institute of Creative Technologies was utilized as the raw data source for our generated depression dataset. The DAIC-WOZ depression database consists of clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder (PTSD) [18]. The length of interviews ranges from 7 to 33 minutes with an average of 16 minutes. DAIC stands for Distress Analysis Interview Corpus, a large corpus in which this database originates, and WOZ indicates the interview approach, namely the Wizard-of-Oz method, where participants interact with a virtual interviewer called "Ellie" that in their believe is autonomous, but in reality, is controlled by an unseen human interviewer in the next room. An illustration of the interview is shown in Figure 3.1 according to [18].

Figure 3.1: An illustration of the Wizard-of-Oz interview. The participant interacts with virtual interviewer, Ellie, during the interview.

The DAIC-WOZ depression database [9] contains visual as well as acoustic recordings and transcriptions of 189 participants (ID ranging from 300 to 492), split into a train set (107 participants), a development set (35 participants), and a test set (47 participants). **Important here to remember is that in all experiments in this thesis, the train set and the development set are merged to construct the training dataset (142 participants) and the original test set is kept as test dataset (47 participants) to test the performance of the developed model in this thesis.** For each session, an eight-item Patient Health Questionnaire depression scale (PHQ-8) is provided as Ground Truth (GT), which indicates the severity of depression, and a PHQ-8 Score ≥ 10 implies that the participant is undergoing a major depression (MD) [24].

3.1 Definition of PHQ-8 System

One of the standardized and validated methods for assessing and diagnosing the severity measure for depressive disorders in large clinical studies is the so-called eight-item Patient Health Questionnaire depression scale developed by Kroenke and Spitzer et al. [23]. The PHQ-8 Score consists of 8 of the 9 criteria (also known as PHQ-8 Subscores), on which the DSM-IV diagnosis of depressive disorders is based [13]. These 8 different aspects of depressive criteria are shown in the Table 3.1 in the section of PHQ-8 Subscores according to [24].

Table 3.1: A Patient Health Questionnaire eight-item depression measure (PHQ-8)

Over the last 2 weeks, how often have you been bothered by any of the following problems?	**Not at all**	**Several days**	**More than half the days**	**Nearly every day**
PHQ-8 Subscores				
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself - or that you are a failure	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
PHQ-8 Score				
Total score ___ = ___ + ... + ___ (sum of all PHQ-8 Subscores, 0 - 24)				
PHQ-8 Binary				
Final result ___ = 1 if PHQ-8 score $\geq$ 10 else 0				

Table 3.1 also demonstrates the definition and relationship between each score in the PHQ-8 system. To obtain the PHQ-8 Score, one would be inquired about the number of days in the past 2 weeks one had experienced a particular depressive symptom. Based on the response and the following conversion: 0 to 1 day means "not at all," 2 to 6 days means "several days," 7 to 11 days means "more than half the days," and 12 to 14 days means "nearly every day," the PHQ-8

Subscore for each criterion is acquired by assigning points (0 to 3) to each category, respectively. The results of PHQ-8 Subscores are then summed up to produce a total PHQ-8 Score between 0 to 24 points, from which a binary state of MD is further derived based on the PHQ-8 Score with 10 as threshold. If PHQ-8 Score $\geq$ 10, it results in an outcome of true classification of having MD, otherwise false. The representation of the depression severity at each numerical range in accord with the PHQ-8 Score is shown in the Table 3.2.

Table 3.2: A representation of PHQ-8 Score

PHQ-8 Score	Level of depressive symptoms	State of MD
0 - 4	not significant	No
5 - 9	mild	No
10 - 14	moderate	Yes
15 - 19	moderately severe	Yes
20 - 24	severe	Yes

So far the definition of the PHQ-8 system (GT of DAIC-WOZ database [9]) has been well explained in-depth. The corresponding underlying relationships among there 3 scores are also established, that is PHQ-8 Subscores, PHQ-8 Score, and PHQ-8 Binary, ranging between 0 to 3, 0 to 24, and 0 / 1, respectively. Hence, it is conspicuous that 3 different prediction scores can be chosen as the output format of the developed depression estimation architecture and either be considered as a classification predictive modeling problem or a regression predictive modeling problem. A classification head provides an advantage of exact prediction by predicting a discrete class label, which resembles the way PHQ-8 structures, whereas a regression head provides an advantage of minimizing the error in decimal places by predicting a continuous quantity. Therefore, several different variations of prediction for such supervised learning task based on DAIC-WOZ database [9] can be found in the previous automatic depression estimation works. Williamson et al. [38] and Gong et al. [16] train their model with regression head by minimizing the RMSE to successfully predict the PHQ-8 Score and further derive the final binary state of MD through cut-off point. Ma et al. [30] and Bailey et al.[4] regard depression detection as a classification problem and solely predict the binary result of MDD of a participant, which is also investigated in other studies [25, 31, 40]. Alhanai et al. [3] and Valstar et al. [37] design 2 models with 2 different output head, one with classification head to model PHQ-8 Binary outcomes and the other with regression head for multi-class outcomes of PHQ-8 Score. Similar to that, Dham et al. [12] also develop 2 models for classification and regression approach. However, instead of predicting PHQ-8 Score, PHQ-8 Subscores were predicted, and the results of final PHQ-8 Score as well as PHQ-8 Binary are calculated according to the definition. More recently, Haque et al. [19], Song et al. [33], and Lin et al. [26] deploy a specific criterion function during the training process to fuse the cross entropy loss and the loss of depression severity assessment since their designed model output with 2 branches, namely a depression classifier for PHQ-8 Binary and a PHQ regression model for PHQ-8 Score.

In this study, in accord with the way PHQ-8 system structures and the consideration of depres-

sion estimation as a classification task, **a classification head, predicting PHQ-8 Subscores, is predominantly exploited in all of the experiments.** PHQ-8 Score as well as PHQ-8 Binary are then derived through the definition, resembling the method in [12]. However, to prove that PHQ-8 Subscores are most informative and do provide the best performance, comparative research is also conducted in chapter 5, which indeed justify this thought. For more details and results about the comparative research, please refer to the Chapter 5.

3.2 Data Understanding

As mentioned above, the DAIC-WOZ [9] depression database contains 189 sessions of interview recordings in total, meaning that 189 interviewees have participated. During the interview, the virtual interviewer, Ellie, will inquire each participant with a subset of possible queries, included direct questions, e.g. "How have you been feeling lately", "Have you ever been diagnosed with PTSD", and response with dialogic feedback, e.g. "I see", "Cool", "That's great." For each session, the recorded data has been transcribed and annotated for a variety of features, which can be categorized into 3 main types, that is visual features, audio features, and text features. In the following parts these features will be discussed in detail and for ethical reasons, no raw video is made available according to [37].

Visual Features: Both low-level and high-level visual features extracted based on *OpenFace* framework [5] and *FACET* software [27] are provided in DAIC-WOZ depression database [9] and listed below. The sampling rate of all visual data is $30\,\mathrm{Hz}$:

- 2D facial landmarks: 68 facial key points in 2D pixel coordinates estimated from video.
- 3D facial landmarks: 3D coordinates of 68 facial key points estimated from video. The points are in millimeters (mm) in world coordinate space, with camera begin at (0, 0, 0).
- Gaze direction: Gaze direction of both eyes in both world coordinate space and head coordinate space
- Head pose: 3D head position and rotation
- HOG: Histogram of oriented gradients on the aligned 112×112 area of the face on the image [10, 28]
- AUs: 17 action units based on facial action coding system (FACS)

In this thesis, the 3D facial landmarks and the gaze direction of 189 individuals are combined and utilized as our visual features to provide micro-facial expression changes for depression estimation. Since the sampling rate of visual data is $30\,\mathrm{Hz}$, the model could be trained to have the capacity of observing microexpression at the millisecond (ms) level. An illustration of micro-facial expression changes based on 3D facial landmarks is shown in the Figure 3.2. It is worth noting that the serial facial landmarks shown here have been speeded up 5 times faster than

normal, meaning that in the visual dataset, the microexpression is much more subtle. Another interesting phenomenon, which could be noticed here, is that most of the gaze directions are pointing down, implying a certain degree of nervousness, sadness, and gloom of the participant. Indeed, this is the clip where the participant answers yes to the question whether he has ever been diagnosed with PTSD or not.

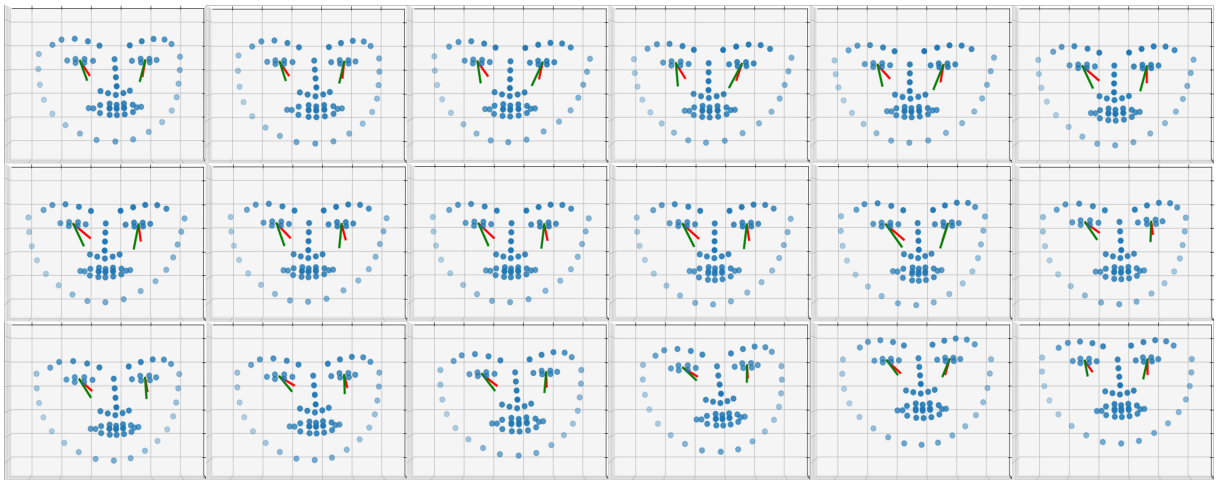


Figure 3.2: An example of Micro-facial expression. The changes of microexpression within 3 s is shown at here(top-left to bottom-right). The red and green marks extending from the eyes visualize the gaze direction in world coordinate and head coordinate, respectively.

Audio Features: The audio data provided in the DAIC-WOZ database [9] consists of a raw audio file, a COVAREP file with pre-extracted audio features, and a formant file, which are recorded over the entire interview. Details of each file are explained below:

- Raw audio file: Recording of the raw audio signal of the whole interview with a head-mounted microphone at a sampling rate of 16 kHz. It was processed so as to possibly only record the voice of the participant. Hence, the voice of the virtual interviewer was silence-suppressed. However, it still contains small amounts of bleed-over of the virtual interviewer.
- COVAREP file: Pre-extracted audio features utilizing the COVAREP(v1.3.2), a freely available open source Matlab and Octave toolbox for speech analyses [11], over the entire recording at every 10 ms, i.e. sampling rate 100 Hz. These involve prosodic features such as fundamental frequency (F0) and voicing (VUV), spectral features including mel cepstral coefficients (MCEP0-24) and harmonic model and phase distortion mean (HMPDM0-24), and energy- or voice quality-related features, for instance, normalized amplitude quotient (NAQ), quasi open quotient (QOQ), maxima dispersion quotient (MDQ), etc.
- Formant file: Containing the first 5 formants, which are the vocal tract resonance frequencies tracked throughout the interview.

In this work, instead of using pre-extracted features in the COVAREP file or formant file, the raw audio file of each individual was preferred and utilized to provide insights into acoustic features

as the model is expected to have the capability of estimating the severity of depression based on audio signal without too many feature engineering techniques, which makes the automatic multi-modal depression estimation feasible for the purpose of serving the public. Imagine how easy it would be for the user to just record their voice and the outcome of MDD will be analyzed directly without any prerequisite of grasping different concepts of higher-level audio features, i.e. MCEP0-24, HMPDM0-24, NAQ, MDQ, etc. A clip of raw audio signal waveform corresponding to the transcript sample shown in the Table 3.3 is illustrated in the Figure 3.3. High-value amplitude areas indicate that the participant is responding to the queries from Ellie, whereas low-value amplitude areas demonstrate the above-mentioned voice suppression while Ellie is speaking.

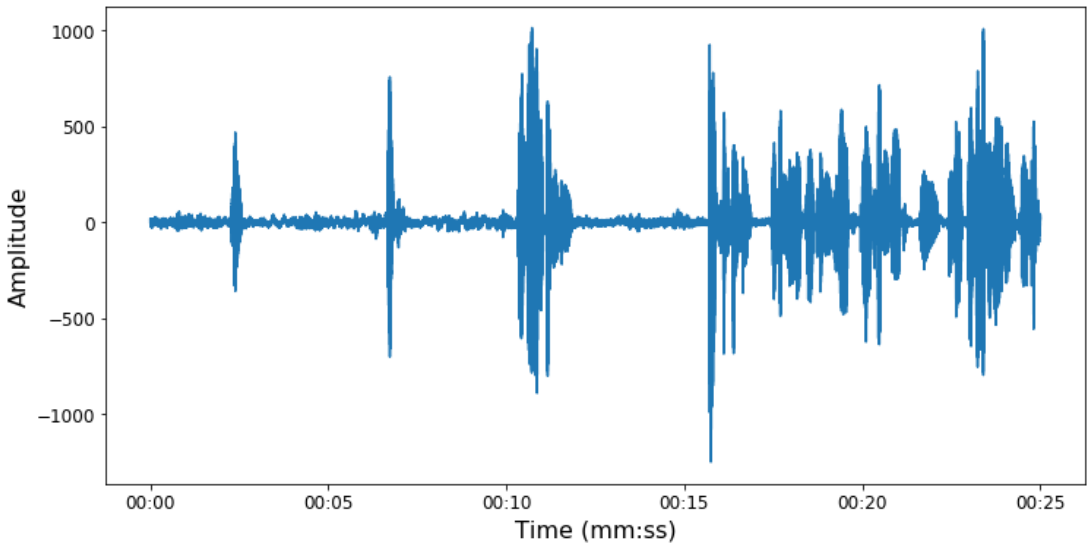


Figure 3.3: An example of Raw audio waveform. Audio signal of Participant 321 beginning from $6'06''$ to $6'31''$.

Text Features: Extensive questionnaire queries of the virtual interviewer as well as verbal responses of the participant are transcribed into text recorded in the transcript file and annotated with the name of the speaker. A sample transcript file is shown in the Table 3.3. In addition, the timestamps of each response are also recorded in a unit of second (s). The transcription conventions according to DAIC-WOZ official documentation are shown below:

- Every word is converted into lower case.
- Incomplete word is annotated with the intended word, followed by a comment with the part that was actually pronounced in angle brackets like: people <peop>.
- Unrecognizable words are indicated as 'xxx'.

In this thesis, the transcript file is not only exploited to contribute textual features to the model, but also the timestamps, which play a critical role in data preprocessing. However, further steps still need to be implemented in order to turn text into an applicable data source since computers can't process words but solely numbers. More details will be illustrated in the Section 3.4.

Table 3.3: An overview of selected samples from the transcript file which are leveraged as text data

Start time in s	Stop time in s	Speaker	Value
...	...	...	...
366.615	368.535	Ellie	have you ever been diagnosed with ptsd
369.375	369.905	Participant	yes
370.960	372.740	Ellie	how long ago were you diagnosed
373.160	374.750	Participant	um it was over five years ago
376.270	377.530	Ellie	what got you to seek help
378.750	380.010	Participant	i couldn't function
380.580	388.990	Participant	i couldn't sleep i couldn't ...
390.350	391.120	Participant	shut down
...	...	...	...

3.3 Potential Problems and Solutions for Dataset

Although the DAIC-WOZ depression database [9] abounds in various data types and data features, which, to a large extent, benefits numerous research for automatic depression estimation with different data-driven approaches, it has been well reported that the DAIC-WOZ depression database [9] contains assorted errors and noises, which will not only cause potential difficulties during the model training process but also sabotage the model performance, which, in the worst case, will mislead the model's attention, leading to completely irrelevant and inapplicable results. Furthermore, it can also be seen from the Section 3.2 that one of the potential problems of insufficient data exists as the DAIC-WOZ depression database [9] consists of only 189 participants. Another potential problem is the imbalanced dataset, involving the imbalance of MD as well as gender imbalance, particularly shown in acoustic features. These potential problems will cause not only failed to fully train a generalized model but also biases. Therefore, in the following subsections, these phenomena as well as solutions for all the above-mentioned problems will be discussed.

3.3.1 Error correction

A list of known errors existing in the DAIC-WOZ depression database [9] is shown below after carefully examining the database:

- GT labeling error: The PHQ-8 Score of participant 409 is 10 but the given PHQ-8 Binary was 0 instead of 1, which contradicts the definition of the PHQ-8 system.
- Missing value error: A value in PHQ-8 Subscores of participant 319 isn't present, which causes *'NaN'* (not a number) problem during the training.
- Data format error: In the development set, the data format of some features of particular participants is different than others. For example, instead of float type, the coordinate of

the facial key points of participant 367 was given in string type. Moreover, since some samplings failed during the interview, those values are assigned to *'-1.#IND'* in string format, meaning not a number. These assigned values have to be replaced in order to process.

- Missing transcript: The transcriptions of the virtual interviewer are missing in some interviews.

To solve the labeling error, missing value error, a pre-check function is created to automatically double-check all the PHQ-8 ground truth of each individual by inspecting the availability of all PHQ-8 Subscores and summarizing them together to form the expected PHQ-8 Score, which will be converted to the expected PHQ-8 Binary based on the definition. These two expected values will then be compared with the provided GT of PHQ-8 Score and PHQ-8 Binary. If only one of the eight values in the PHQ-8 Subscores of a participant is missing, this lost value will be automatically calculated according to the given PHQ-8 Score. However, if there are multiple missing values, then the process will be terminated and the ID of this problem participant will be printed out as there isn't any solution for this. Fortunately, this multiple missing values problem doesn't exist. Another task of this pre-check function is to ensure the correct data format for each loaded feature. All of the loaded data will be converted into a float format. For non-numeric values such as *'NaN'* or *'-1.#IND'* will be set to 0. As for the missing transcription of the interviewer, since only the features of the participant were focused on and extracted in this work, meaning that only the sentences and timestamps of the participant are desired, it doesn't lead to any complication and thus is ignored.

3.3.2 Noise reduction

Similar to errors, the acknowledged noises in the DAIC-WOZ depression database [9] are listed below:

- Acoustic noise: There are usually strong noises at the beginning and end of the audio recording because the agent is supporting the interviewee to settle down while the audio recording has already started.
- Bleed-over of the voice of virtual interviewer: Small amount of Ellie's voice is still recorded in spite of silence suppression shown in figure 3.4.
- Interruption during the interview: Some interviews involve long or short interruptions caused by technical issues or ringing of the cellphone, which should be removed.
- Irrelevant interaction: Each recording contains interactions between the participant and the researcher prior to the beginning of the interview which needs to be removed.

As disclosed above, most of the concern about noises relates to acoustic data. Hence, in this subsection, an example of the process of audio data cleaning will be well explained in-depth and an illustration is demonstrated in the Figure 3.4. Three different audio signal waveforms

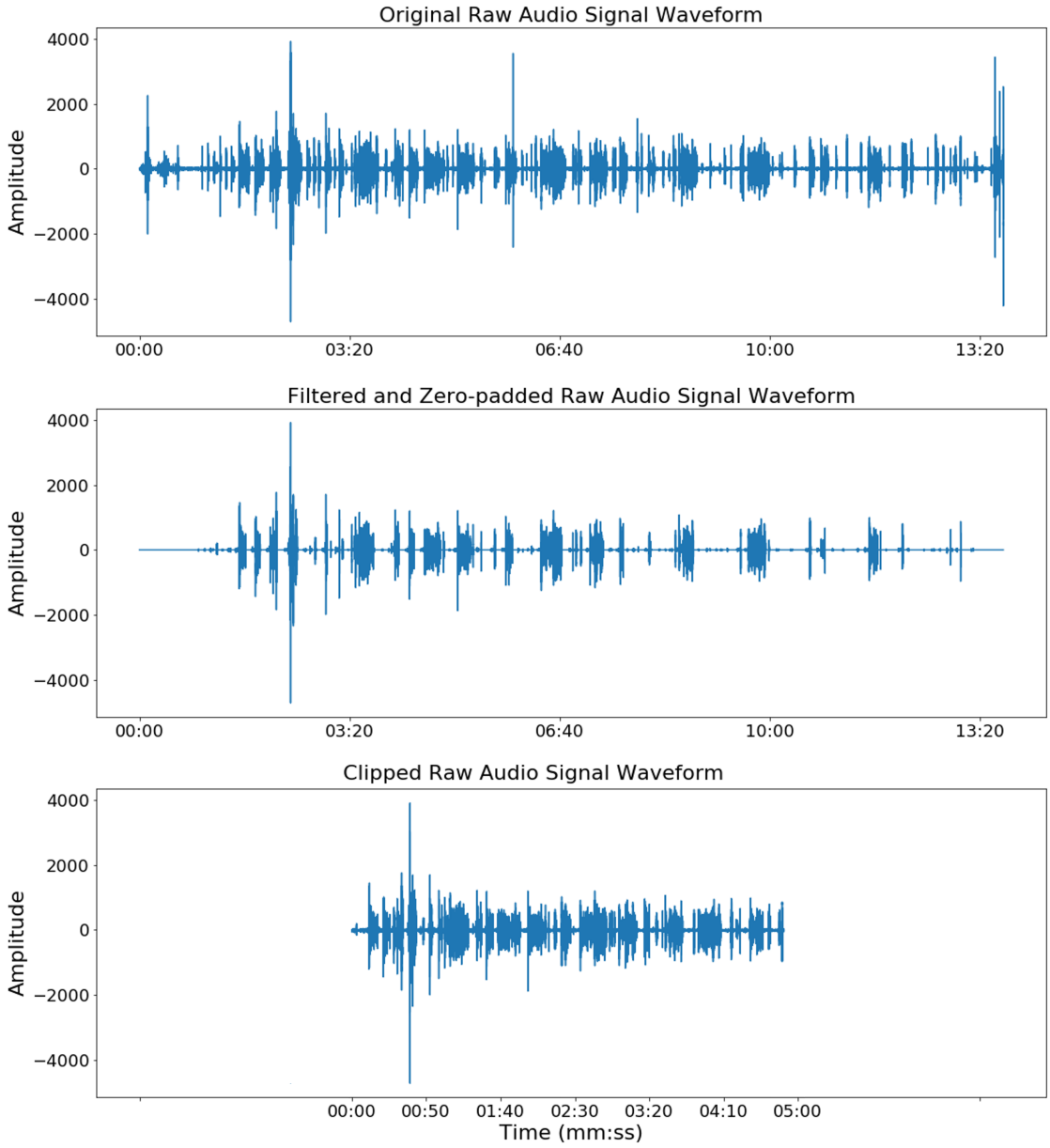


Figure 3.4: An illustration of noise reduction of audio signal, which includes the illustration of the original audio signal, interim result after filtering, and the final clipped audio signal exploited in this work.

are shown here, namely "Original Raw Audio Signal Waveform", "Filtered and Zero-padded Raw Audio Signal Waveform", and "Clipped Raw Audio Signal Waveform", ordering from top to bottom. As their name implies, distinct methods have been applied respectively to clean the data. The raw signal waveform in the first plot is the initial and unprocessed audio data and the filtered signal waveform in the second plot is the interim result after filtering out all of the noises as well as voices besides the participant in alignment with the audio length of the raw signal waveform. By comparing these two, it is conspicuous that the raw signal consists of a huge amount of acoustic noises, the bleed-over voice of the virtual interviewer, etc. Therefore, it

is then cropped and clipped based on the timestamps provided in the transcription and includes only the acoustic features of the participant. This final outcome is illustrated in the third plot of the Figure 3.4, which is also the ultimate audio signal utilized in this thesis later in the Section 3.4, "Preprocessing".

3.3.3 Handling of an imbalanced dataset

A major challenge in training a shallow or deep depression estimation model with the DAIC-WOZ database [9] lies in the unequal distribution of the dataset, including an uneven sample of depressed and non-depressed participants as well as gender imbalance, particularly appeared in acoustic features. It has been widely reported that imbalanced classes in a dataset will greatly affect the performance of the ML model. Moreover, many current benchmarks [4, 16, 30, 3] have shown great adversity of undergoing data imbalance among different levels of depression, which incurs a large bias in the predicted results. Hence, several techniques have been developed to solve this problem. Ma et al. [30] applied the random sampling technique to the non-depressed class to randomly crop for each subject to match the number of clips in the depressed class, on which no particular operation has been conducted. Gong et al. [16] performed random-oversampling by simply duplicating samples to make the number of samples for each PHQ-8 Score roughly equal.

In this subsection, the illustrations of imbalanced phenomena and the approaches to coping with such problems in this thesis will be shown and discussed in detail.

Imbalance between depressed and non-depressed participant: DAIC-WOZ depression database [9] involves the recording of 189 participants, split into train set, development set, and test set. As aforementioned, in this thesis, our generated training dataset (142 participants included) is based on the combination of the train set and development set for the training process and the the test set is only composed of the provided test set in this dataset (47 participants included). Since only the training dataset is related to the model training itself, only the distribution in the training dataset will be discussed here. However, the distribution in the test dataset also resembles this result in the training dataset. An overview of the distribution of depressed (D) and non-depressed (ND) participants, as well as female (F) and male (M) participants, are shown in the Table 3.4.

Table 3.4: An overview of the distributions of PHQ-8 Binary Score regarding gender and major depression

Training	Female (F)	Male (M)	Total F+M
ND	39 (28%)	60 (42%)	99 (70%)
D	24 (17%)	19 (13%)	43 (30%)
Total ND+D	63 (45%)	79 (55%)	142 (100%)

On first look at the last column, it can be noticed that the proportion of depressed to non-depressed participants is about 3 to 7, meaning that only 30% of the participants are being

classified as depressed, whereas non-depressed participants constitute about 70% of the participant, which is about 2.3 times larger than that of depressed ones. Furthermore, between each gender, the distribution of PHQ-8 Binary in male participants shows a greater imbalanced phenomenon than in the female. To further dive deep into PHQ-8 Score as well as PHQ-8 Subscores, bar graphs of both are shown in the Figure 3.5. One can tell that no matter which subclass in PHQ-8 Subscores or PHQ-8 Score, the tendency slants to the right side, meaning a strong bias toward non-depressed class, thus making this database not reliable by directly adopting it for model learning.

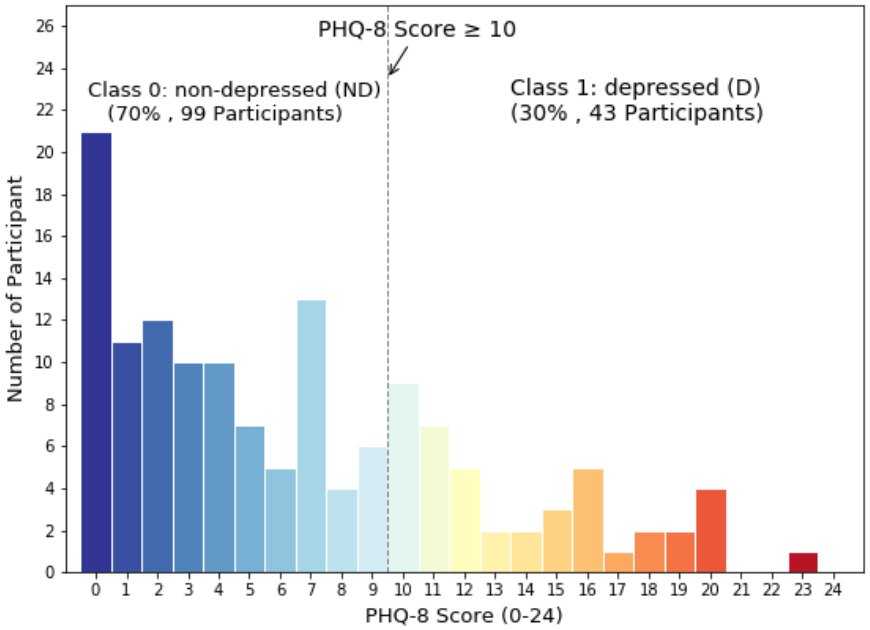
To solve this, a "weighted random sampler" in PyTorch is utilized for the data loader to equally load the data from each class. However, instead of loading each PHQ-8 subclass (0-4) equally, which is the predicted score for our 8 classification heads mentioned previously, the PHQ-8 classes corresponding to the PHQ-8 Score, ranging from 0 to 24, are loaded equally since PHQ-8 Subscores are fixed to each participant and there isn't any other way to equalize the number of subclasses while loading the batches based on either the participants or clips of the participant. Moreover, a dynamic weighted loss function is also applied to minimize this imbalanced phenomenon rather than a loss function without any weight or a static weighted loss function. Due to the fact that the subclasses can't be loaded equally between each batch, only the distribution of each subclass per batch can be derived. Therefore, the weight for the weighted loss function is dynamically calculated according to the number of each subclass of each batch throughout the training.

Gender imbalance, particularly in acoustic feature: Another imbalance one can observe from the table 3.4 is the gender imbalance. In total, female constitutes about 42% of the participants (equal to 68 people), where 39 female participants are non-depressed and 24 female participants are having MD, while male constitutes about 55% of the participants (equal to 79 people), where 60 male participants are non-depressed and only 19 male participants are having MD. There is a 10% difference shown in the number of female and male participants. This imbalance affects acoustic features particularly strong as one could imagine that it is hard to tell just based on the facial key points shown in the Figure 3.2 or text shown in the Table 3.3 whether this person is a man or a woman, whereas it is conspicuous just by listening to the voice of the recording or observing the fundamental frequency (F0) in the spectrogram.

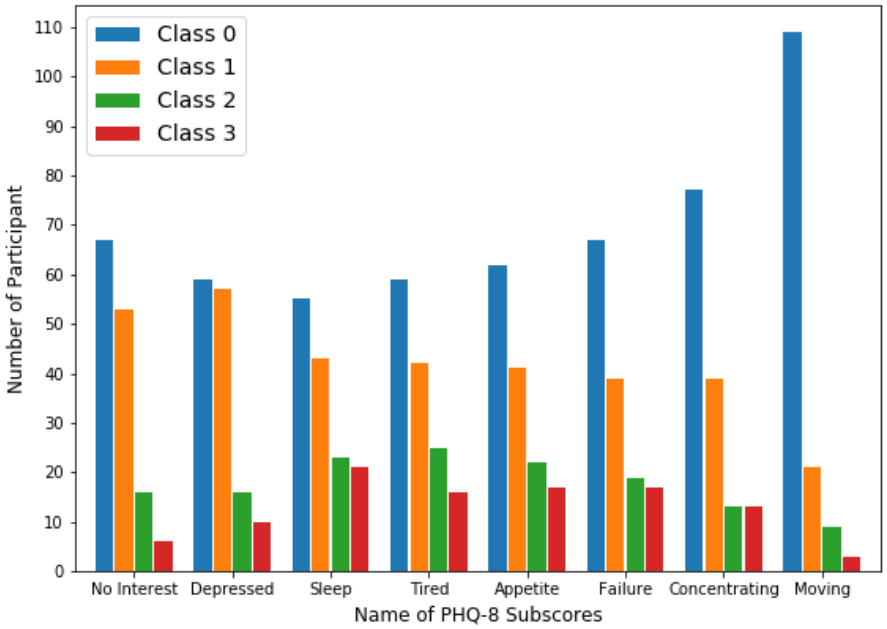
Therefore, to solve this specific problem for audio and ameliorate the phenomenon of acoustic bias or so-called gender bias, an online software tool [1] is exploited to convert each voice of the participants in the recording to the contrary gender for gender balancing, and a new audio dataset is then generated, specifically to train the backbone of the audio branch in our multi-modal model to have the better as well as non-biased capability of extracting depressive characteristic of MD.

3.3.4 Overcoming of a small-scale dataset

The other potential problem of the DAIC-WOZ depression database [9] is the scale of the database. There are only 142 sessions in the training dataset, i.e. interviews of 142 participants, which is a relatively little number of samples compared with the complexity of the depression



(a) Distribution of PHQ-8 Score



(b) Distribution of PHQ-8 Subscores

Figure 3.5: An overview of the PHQ-8 distribution. Distribution of PHQ-8 Score (a) and PHQ-8 Subscores (b).

estimation task. This small-scale dataset not only leads to a hard time to train a representative model but also incurs failures of the generalization ability of the model. This means that the model can encounter at least the following issues: overfitting, underfitting, outliers, sampling bias, missing values, etc. Hence, the sliding window technique is applied to segment the interview

into N overlapped clips and increase the dataset size, with a window size of $60\,\mathrm{s}$ and an overlap size of $10\,\mathrm{s}$.

The final outcome is illustrated in the figure 3.6. Before applying the sliding window technique, there are only 142 and 47 participants in the training dataset and test dataset respectively. After that, the scale of the dataset has been expanded tenfold, that is there are around 13 hundred clips to train the model and around 5 hundred clips to test the model's accuracy. With the help of this technique, our model shows a significant performance improvement and stability, which will be demonstrated more in detail in the Chapter 5.

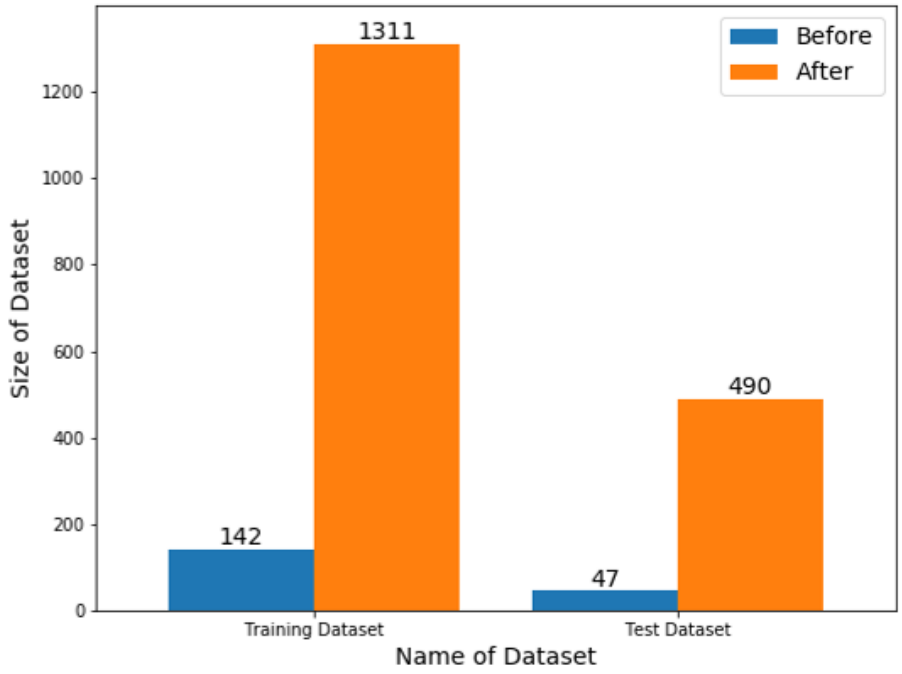


Figure 3.6: Augmentation of the dataset. The result before and after applying sliding window technique to clip and increase the scale of the dataset.

3.4 Preprocessing

So far the DAIC-WOZ database [9] has been introduced in-depth, along with potential problems such as errors and noises as well as our proposed solutions. In this section, all these techniques are going to be combined into a preprocessing framework for each data type designed to remove these errors to provide a cleaner dataset, which is also in a uniform data format, to utilize for the model training and testing.

3.4.1 Text Data

As illustrated in the Table 3.3, the contents of the transcription record the transcribed conversation between the virtual interviewer and the participant together with the timestamps as well as the speaker's name. However, since only the features of the participant are interested in this thesis,

only the features from each interviewee are extracted, namely the sentences of each participant. Furthermore, the start-stop time pair of each sentence of the participant has also been extracted as it is essential to filter and ensure that the timeline between each input data domain is roughly aligned with each other so that the late fusion will not lose its meaning. An illustration of text data preprocessing is demonstrated in the figure 3.7. In addition to sentence extraction, a further step has to be conducted to transform and represent text in a language that a computer could be able to process and understand, i.e. sentence embedding. Instead of utilizing word embedding, which deals with individual words, sentence embedding is chosen here as it not only retrieves information more efficient compared with word embedding by converting the whole sentence into a vector, but also helps the computer in understanding the context, intention, and other nuances in the entire text.

In this work, the pre-trained model of universal sentence encoder large from Google [17, 6] is exploited as our sentence embedding model. The model has a Transformer encoder-like architecture and is trained on a variety of data sources and a variety of tasks to dynamically accommodate a wide variety of natural language understanding tasks according to [17]. It encodes variable-length English text as well as sentences and outputs a 512-dimensional vector that can be used for text classification, semantic similarity, clustering, and other natural language tasks. Of course, there is also a possibility of training our sentence embedding model, which could even boost the performance. However, in our opinion, this would lead to a problem that the model would not have a generalization ability to understand different texts besides from the DAIC-WOZ depression database [9] or overfit with the text as the problem mentioned before. That is why the universal sentence encoder from Google has been chosen.

3.4.2 Visual Data

As illustrated in the Figure 3.2, the visual data consists of the micro-facial expression of facial key points and gaze direction. However, the raw data of both in the DAIC-WOZ database [9] are in fact separate, unnormalized, and need to be reformatted and cropped out the irrelevant parts. Therefore, this preprocessing for the micro-facial expression has been developed. An example of raw visual data is demonstrated in the table 3.5, including 2 sample frames of facial key points in the table 3.5a and gaze direction vectors in the table 3.5b.

As you could tell from the the Table 3.5, either facial key points or gaze direction are given in 2D format instead of 3D format, even though there are both 3D data features. 68 3D facial key points are annotated from 0 to 67 with X, Y, Z corresponding to its axis and 4 3D gaze direction contains the gaze direction vectors of both eyes in both world and head coordinates, which is marked with "h". 0 represents the right eye, whereas 1 represents the left eye. Hence, the first step of the preprocessing is to reformat them into the standard 3D coordinate format, where the last dimension represents the three axes, namely the x-axis, y-axis, and z-axis. For 68 3D facial key points, they are transformed from $(T \times 204) \longrightarrow (T \times 68 \times 3)$; for 4 gaze direction vectors, they are transformed from $(T \times 12) \longrightarrow (T \times 4 \times 3)$, where T indicates the number of the frame. Secondly, one can notice that the range of the axes in facial key points varies a lot, i.e. the x-axis from -67 to -142 but the z-axis from 505 to 565, which could potentially lead to a problem of

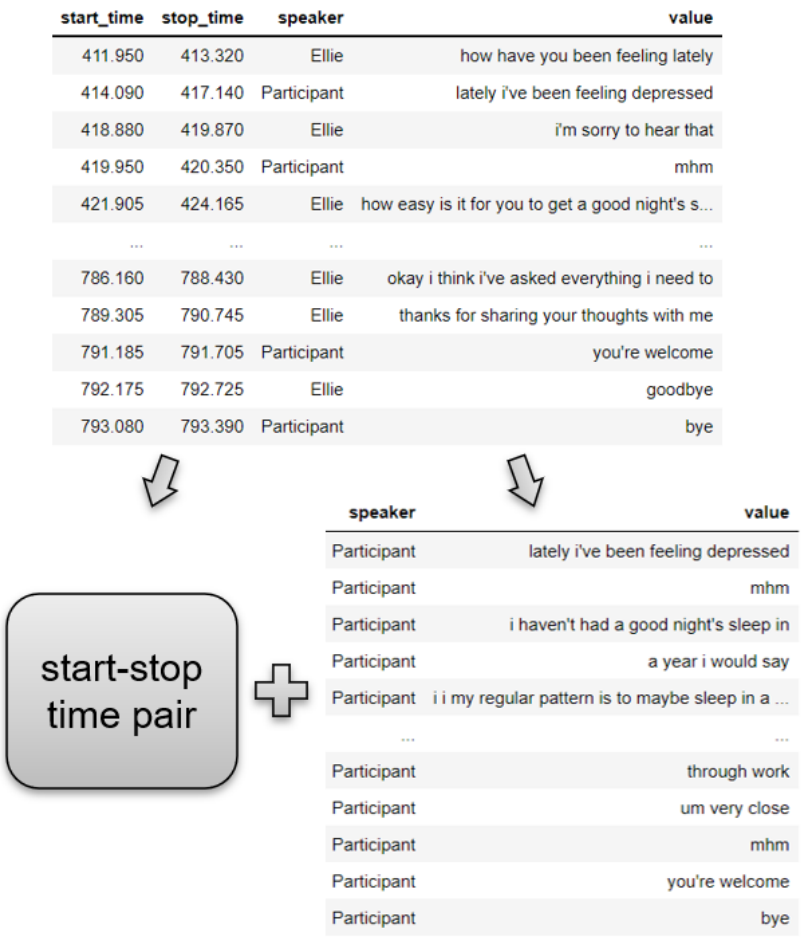


Figure 3.7: Preprocessing of text data. Sentences and timestamps of the participants are being extracted.

Table 3.5: Example of raw visual data. Coordinate of 68 3D facial key points in 3.5a and 3D gaze direction vectors in 3.5b

Frame	X0	X1	…	X67	Y0	Y1	…	Y67	Z0	Z1	…	Z67
11730	-142.2	-142.5	…	-67.6	-72.1	-51.5	…	4.2	565.9	569.6	…	505.5
11731	-141.4	-141.9	…	-67.5	-71.1	-50.7	…	4.6	565.4	568.7	…	504.6

a . 68 3D facial key points

Frame	x_0	y_0	z_0	x_1	y_1	z_1	x_h0	y_h0	z_h0	x_h1	y_h1	z_h1
11730	0.17	0.28	-0.95	-0.03	0.38	-0.92	0.08	0.38	-0.92	-0.11	0.49	-0.86
11731	0.27	0.32	-0.90	-0.05	0.39	-0.91	0.19	0.42	-0.88	-0.12	0.50	-0.85

b . 4 3D gaze direction vectors

bias between axes. Therefore, the facial key points are normalized with the following equation:

$$\mathbf{X}' = a + \frac{(\mathbf{X} - \mathbf{X}_{min})(b - a)}{(\mathbf{X}_{max} - \mathbf{X}_{min})}, \tag{3.1}$$

where the input and the output: $\mathbf{x}, \mathbf{x}' \mapsto \mathbb{R}^3$ and $a = 0, b = 1$ as the facial key points are normalized to range: 0 - 1. As for the gaze direction, no normalization is required since the given vectors either in world coordinate or head coordinate are unit vectors already. After the normalization, both of them are combined into a larger matrix with a dimension of $(T \times 72 \times 3)$. Finally, the parts where Ellie speaks as well as irrelevant interactions are cropped out based on the start-stop time pair extracted from the text preprocessing part shown in the Figure 3.7. Now, a clean, as well as normalized visual data of micro-facial expression as illustrated in the Figure 3.2, has been prepared for direct utilization of the model learning.

3.4.3 Audio Data

In the previous Subsection 3.3.2, it has been well discussed why and how the original raw audio signal is processed to achieve the final clipped raw audio signal shown in the third diagram of the Figure 3.4. However, instead of utilizing the clipped raw audio signal, a more informative and effective audio data format is preferred because the raw audio signal itself still consists of plenty of redundant segments.

The first idea is to transform the clipped raw audio signal into the log-spectrogram by applying STFT and nonlinear transformation. The outcome is illustrated in the Figure 3.8 with time (s) in x-axis and frequency (Hz) in y-axis. To observe the difference between both gender and having major depression, the following spectrograms are included: the woman without MD, the woman with MD, the man with MD, and the man without MD. At first glance, one could notice that for each spectrogram, there is always a particular light band crossing the whole spectrogram. For the

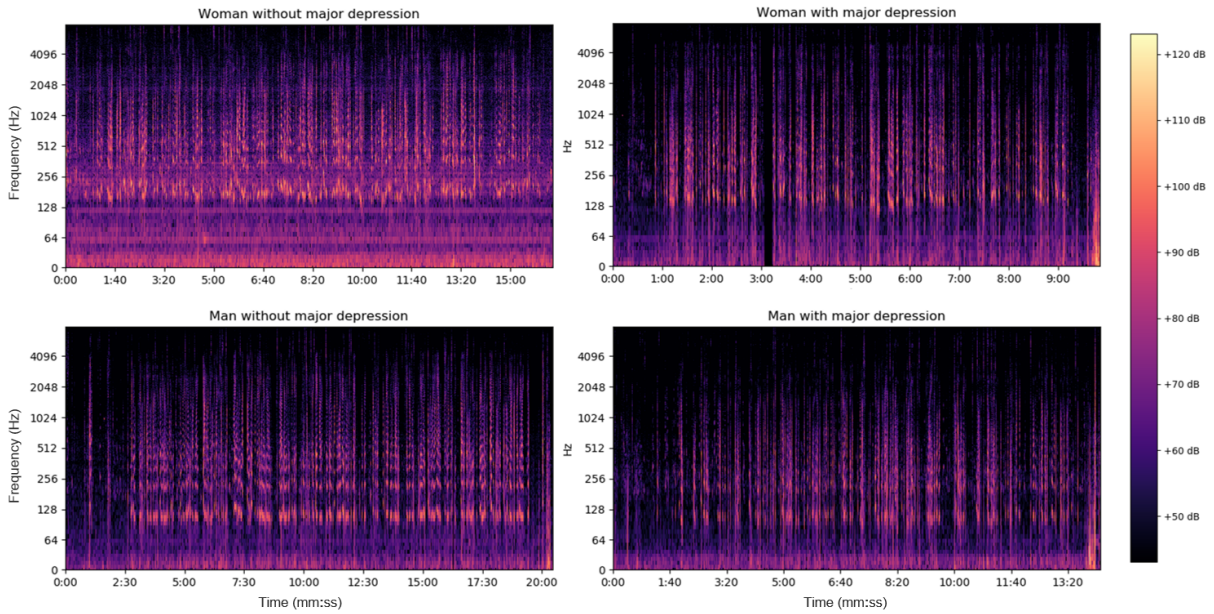


Figure 3.8: A comparison of log-spectrogram.

female spectrogram, it is around 210 Hz and for the male spectrogram around 120 Hz. In fact, these particular frequencies differed in gender are the so-called "fundamental frequency (F0)", whose values seen here are in accord with the research from Traunmüller et al. [36]. Moreover, it is observable that the spectrogram from the female with MD resembles the spectrogram from the male without MD, which, in our opinion, could potentially cause confusion to the model and thus harm the performance. Therefore, in order to increase the difference and avoid this problem, the second approach is come up, i.e. log-mel spectrogram.

Log-mel spectrogram differs from spectrogram by converting the frequency scale in spectrogram to the so-called mel-scale, which is a more human perception-like frequency scale designed by Stevens et al. [34]. As you can imagine, it is not a huge challenge for us humans to distinguish the voice of both genders. Hence, by exploiting this technique, there is no doubt that the subtle differences in Figure 3.8, especially between the female with MD and the male without MD, will rise and the complication will be ameliorated to the greatest extent possible. Illustrations of the converted log-mel spectrograms of identical individuals as in spectrograms are shown in the Figure 3.9. This time, by perceiving the apparent difference of F0 from both genders and comparing the log-mel spectrogram between the female with MD and the male without MD, it can be concluded that this approach brings great success in solving this issue. Moreover, based on the observation, the color of the log-mel spectrogram of the participants with MD is darker than the participants without MD, indicating that they usually speak quietly and less energetically, which could be a potential characteristic of having major depression.

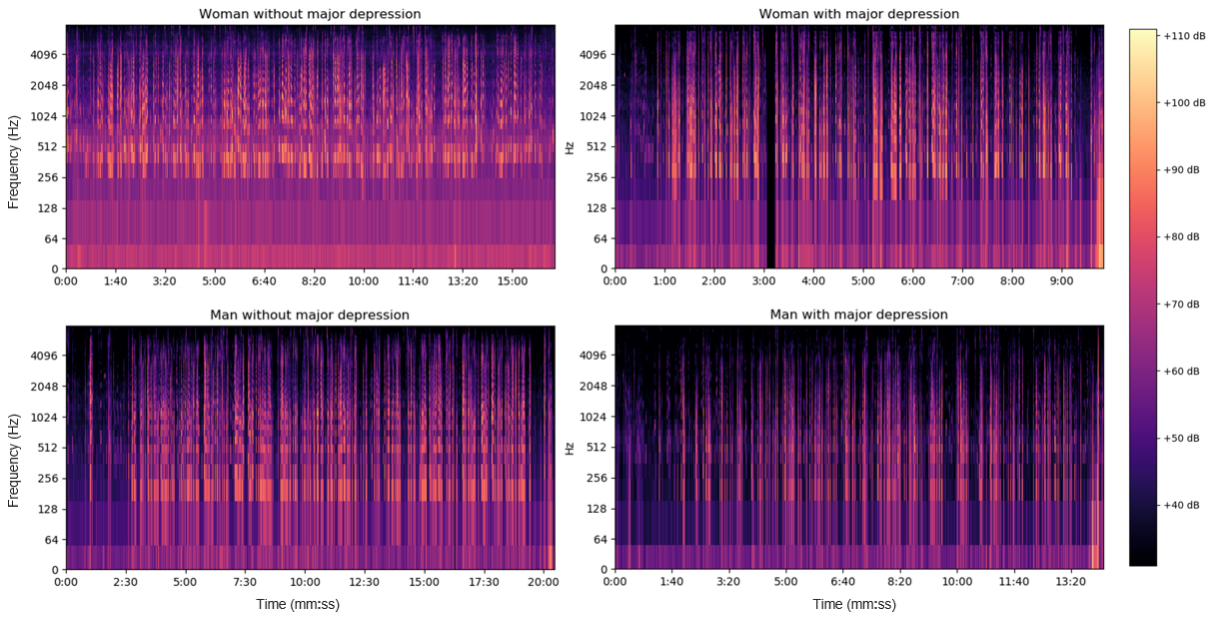


Figure 3.9: A comparison of log-mel spectrogram.

Finally, after taking all the above-mentioned perspectives into consideration, the preprocessing framework for the acoustic features could be summarized in the following 4 steps:

1. Cropping out all the parts from Ellie and noises form the original raw audio signal to form the clipped raw audio signal based on start-stop time pair extracted from the transcription.

2. Applying STFT to extract spectrogram from the clipped raw audio signal, and transforming it to mel spectrogram by converting the frequency scale to mel-scale. Here for STFT, Hann function was used with a window size of 2048 and hop size of 533, which is derived from the sampling rate of audio data and visual data, namely $\frac{\text{acoustic sampling rate}}{\text{visual sampling rate}} = \frac{16\,\text{kHz}}{30\,\text{Hz}}$. The number of the mel band for mel-scale is 80.

3. Converting the mel-spectrogram with linear transformation to log-mel spectrogram.

4. Standardizing the log-mel spectrogram with the following equation:

$$\mathbf{X}' = \frac{\mathbf{X} - \mu}{\sigma}, \tag{3.2}$$

where μ: mean of input $\mathbf{X}$ and σ: standard deviation of input $\mathbf{X}$.

3.5 Dataset Generation

Throughout this work, three different datasets have been generated for the various purposes in the experiment, i.e. raw dataset, clipped dataset, and audio dataset. By summing up the points as well as processes mentioned in the Section 3.3 and 3.4, three different scripts are created to automatically generate these datasets, which also indicates that three data loaders are created to load each dataset. In this section, details, as well as the objectives of each dataset, will be explained.

3.5.1 Raw Dataset

The raw dataset is the original DAIC-WOZ depression database [9] without any above-mentioned feature engineering techniques such as data clipping, gender balancing, sampler, etc. For this dataset, the data loader loads the data based on the participant, meaning that for a batch size of 10, the whole recording of 10 different individuals including visual data of micro-facial expression, audio data of log-mel spectrogram, and sentence embedding of transcription will be loaded into the model for model learning and the result will then be predicted. At first glance, this seems to be logical because this is exactly how the interviewers perceive and diagnose whether a patient has MD or not. They interrogate the participant with several questions, meanwhile, observing his or her reaction as well as the expression. After the whole interview is finished, the decision is then made. However, for computers, one problem that could directly be realized is that this loaded data size is too enormous, leading to several difficulties, e.g. inefficient training process, infinite time to train, running out of memory problems, etc. The interview length is ranging from 7 minutes to 33 minutes with an average of 16 minutes and it costs at least $25\,\text{GB}$ to load each data type. Moreover, since no techniques have been applied to solve and handle the imbalanced and small-scale dataset problems. It is obvious and to anticipate that training the model, either multi-modal or single-modal, based on this raw dataset will fail in all experiments. Indeed, proven by the results from the failure of the experiments in the Section 5.2, in which the model only predicts class 0 (non-depressed) as the result no matter which participant is being

processed, it can be concluded that the raw dataset fails and thus is being abandoned in the work. That is why our focus is turned to the second generated dataset, namely the clipped dataset.

3.5.2 Clipped Dataset

As opposed to the raw dataset, the clipped dataset is the edited original DAIC-WOZ depression database [9] by applying various feature engineering techniques mentioned above in the Section 3.3 and 3.4. Moreover, in order to reduce the computation load while model training, the sliding window technique is utilized to clip the whole interview into several segments with a window size of $60\,\mathrm{s}$ and an overlap size of $10\,\mathrm{s}$. The data loader, in this case, loads the data based on clips instead of participants, implying that the model predicts the depression status of each clip rather than each participant. If an interview of a participant has been split into 10 clips, these 10 clips will be loaded into the model, possibly in different batches since the weighted random sampler is exploited to randomly as well as equally load the clips from each different class, which will then outcome 10 groups of predicted PHQ-8 Subscores for each clip even though there are all from the same participant. The pseudocode of the script to automatically generate the clipped dataset is shown in algorithm 1.

Algorithm 1: Generation of clipped dataset

Input : $dataset\_path$, $output\_path$
Output : A new generated dataset

Initialize new_GT, $window\_size$, $overlap\_size$
$participants$ = ***LoadPaticipants***($dataset\_path$)
for $participant$ *in* $participants$ **do**
  $GT \leftarrow$ extracting the GT of current participant
  $visual\_path$, $audio\_path$, $text\_path \leftarrow$ getting all file paths of each data type
  % Loading all data type
  $visual\_features$, $visual\_sr$ = ***LoadVisual***($visual\_path$)
  $audio\_features$, $audio\_sr$ = ***LoadAudio***($audio\_path$)
  $text\_features$, $time\_pair$ = ***LoadText***($text\_path$)
  % Clipping the data
  $clipped\_visual$ = ***VisualClipping***($visual\_features$, $visual\_sr$, $time\_pair$)
  $clipped\_audio$ = ***AudioClipping***($audio\_features$, $audio\_sr$, $time\_pair$)
  $mel\_spectrogram$ = ***Normalize***(***ConvertMel***($clipped\_audio$, $audio\_sr$, $frame\_size = 2048$, $hop\_size = 533$, $num\_mel\_bands = 80$))
  % Applying sliding window technique and storing the segments
  $num\_segment$ = ***SlidingWindow***($clipped\_visual$, $mel\_spectrogram$, $text\_features$, $visual\_sr$, $window\_size$, $overlap\_size$, $output\_path$)
  $new\_GT \leftarrow$ storing the $num\_segment$ times duplicated GT in $output\_path$
end

After turning to the clipped dataset instead of the raw dataset, great success has been shown in

all experiments. Therefore, the clipped dataset is predominantly utilized in the experiment.

3.5.3 Audio Dataset

The third dataset that is utilized throughout this thesis is the audio dataset. As mentioned before, this audio dataset is specifically generated for solving the gender bias phenomenon in acoustic features. To generate it, the gender balancing technique pointed out in the Subsection 3.3.3 has been deployed by converting the audio recording of a participant to the contrary gender voice. Hence, for each participant, there are now 2 raw audio signals available, namely, the original one and the transgender one. These 2 raw audio signals are then passed through the preprocessing steps listed in the Subsection 3.4.3 to extract the final, normalized log-mel spectrograms, and additionally, similar to the clipped dataset, the sliding window technique is exploited to cut the whole log-mel spectrograms into several clips to reduce the computational load later in the training process.

With such a deployment of the gender balancing technique to address this problem, the impact of the gender bias could be maximally alleviated and the model could have the true competence of estimating the depression severity solely according to the acoustic depression characteristic and ignoring the gender differences. Therefore, this newly generated audio dataset is used particularly to train the audio backbone either in the single modality audio model or in the multimodality model. In fact, proving by the result of the experiments, a significant performance improvement can be observed after training with this dataset.

Chapter 4

Model

Throughout this thesis, several multi-modal models, as well as single-modal models with different feature extractors, have been implemented and tested. The final best model is the so-called Sub-Attentional ConvBiLSTM and the second-best model is the Attentional ConvBiLSTM. These two networks are all attentional multi-modal models implemented with the late fusion of extracted features from three different modalities, i.e. visual, audio, and text. Furthermore, a specific score distribution generation technique called "Multi-path Uncertainty-aware Score Distributions Learning (MUSDL) [35]" for converting each hard-label in ground truth to a soft-label score distribution is applied, along with the Kullback-Leibler (KL) divergence [21] for the calculation of learning loss. As for the optimizer, instead of the first-order optimization function, the second-order optimization function, "Sharpness-Aware Minimization (SAM) [15]", is utilized. SAM is specifically devised and has been proven from the previous works to improve the generalization ability of the model even just training on a small dataset. Therefore, in this chapter, the architecture of both models as well as the attentional fusion layer will be introduced. Similarly, the functionality of MUSDL and SAM regarding the contribution of model performance improvement will be discussed.

4.1 Sub-Attentional ConvBiLSTM

The model architecture of Sub-Attentional ConvBiLSTM is illustrated in the Figure 4.1.

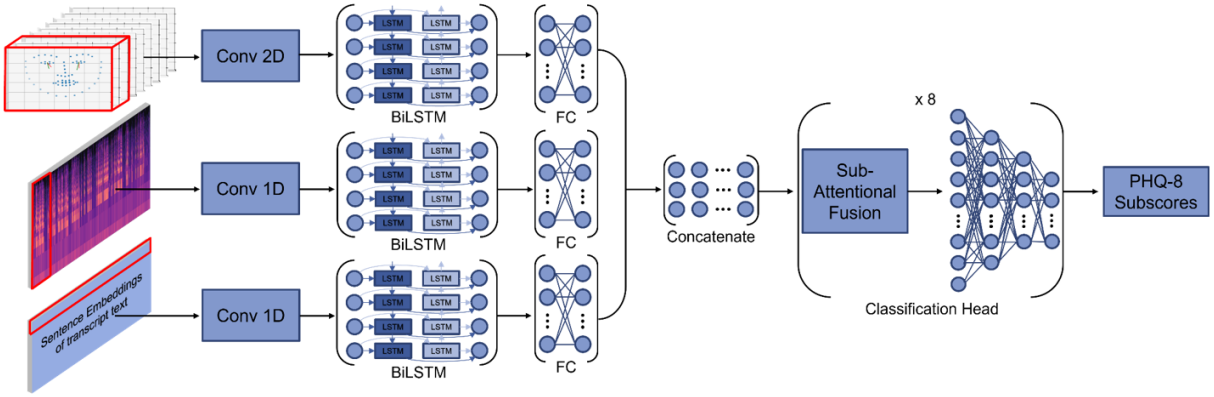


Figure 4.1: An architecture of Sub-Attentional ConvBiLSTM.

By leveraging multi-modal data, deeper underlying depression cues can be encoded. Hence,

three different feature domains have been deployed as model input shown in the Figure 4.1, namely visual input of micro-facial expression, audio input of log-mel spectrogram, and text input of sentence embeddings, which will then be processed by each backbone to extract the higher-level representation of each feature.

The backbone chosen here is based on "DepAudioNet" from Ma et al. [30], which was one of the best models in AVEC 2016 by exploiting a deep learning-based approach with solely acoustic features for depression detection. In general, DepAudioNet is a serial combination of 1D-CNN layers and LSTM layers with a 2-dimensional input format. In this thesis, it has been further improved by transforming LSTM layers into bidirectional LSTM (BiLSTM) layers and enabling the applicability to a 3-dimensional input format for visual input by utilizing 2D-CNN layers. Therefore, it is termed ConvBiLSTM in this work. The detailed configuration of the improved backbone of ConvBiLSTM is summarized in the Table 4.1 and 4.2.

Layer Name	Parameter Settings
Conv2D	Hidden: 256 Kernel: (72, 3) Stride: (1, 1) Pad: (0, 1) Norm: BatchNorm2D Activation: ReLU
MaxPool1D	Kernel: 3 Stride: 3 Pad: 0
Dropout	0.5
BiLSTM	Hidden: 256 Layers: 4 Dropout: 0.5
FC	Output features: 256 Norm: BatchNorm1D Activation: ReLU

Table 4.1: Configuration of Conv2D-BiLSTM.

Layer Name	Parameter Settings
Conv1D	Hidden: 256 Kernel: 3 Stride: 1 Pad: 1 Norm: BatchNorm1D Activation: ReLU
MaxPool1D	Kernel: 3 Stride: 3 Pad: 0
Dropout	0.5
BiLSTM	Hidden: 256 Layers: 4 or 2 Dropout: 0.5
FC	Output features: 256 Norm: BatchNorm1D Activation: ReLU

Table 4.2: Configuration of Conv1D-BiLSTM.

In the audio and text branch, a 1-dimensional CNN layer has been first utilized in the backbone to provide translation-equivariant responses of a low-level feature map, whose kernel size k is 3, indicating that several short-term features are captured at this layer. In contrast to that, a 2-dimensional CNN layer is exploited for the visual branch as visual data has a 3D format instead of a 2D format like audio and text. However, rather than using a (3×3) square-shaped kernel, all 72 key points are perceived as a whole and the kernel slides through the visual data solely along

the time-axis, focusing on extracting temporal changes between each frame and coordinate to provide local attention. Hence, the kernel size is (72×3). Then, batch normalization is performed to regularize the intermediate representation of the features to a standard normal distribution, followed by a nonlinear transformation with the rectified linear unit (ReLU), an activation function defined as the positive part of its argument: $f(x) = max(0, x)$. To further reduce the dimensionality, a max-pooling layer is applied to down-sample the input representation of the feature map by allowing for the assumptions to be made about features contained in the sub-regions binned. A BiLSTM layer together with a FC layer is stacked at the end of the backbone structure, for the objective of harvesting long-range variability in each modality along the time-axis and retrieving effective features from each branch. After all features from each modality are extracted, they are then concatenated in parallel to form a feature map as input to the subsequent late fusion layer.

Considering the superiority of weighting fusion methods over the traditional fusion methods and for more effective deployment of such feature map of muli-modal data, this work presents a novel idea of inserting attentional fusion layers inspired by Dai et al. [8] between the model backbone and output head to realize attentional information interaction between each modality before passing to output head. The major difference between the weighting fusion methods and traditional fusion methods is whether there is an extra training layer or not. Details of the structure, as well as the functionality of the attentional fusion layer, will be discussed in the Section 4.3. In this Sub-Attentional ConvBiLSTM model, 8 different attentional fusion layers are exploited, connected with 8 different output heads, respectively. 8 at here corresponds to the subclass number of the PHQ-8 Subscores as they are our predicted score, That is why this model is named Sub-Attentional ConvBiLSTM. With this structure, each sub-attentional fusion block will be trained to have the competence in focusing on different depression cues from different modalities according to the demand of each subclass, which might have different attention from the concatenated feature map. For output heads, classifiers are used as depression estimation with the PHQ-8 system is considered as a classification problem based on the definition in this thesis. As for the final PHQ-8 Score, the indicator of the severity of depression, as well as PHQ-8 Binary, the binary state of suffering from MD, are further derived based on the definition described in the Section 3.1.

With such network architecture, it is thus expected to not only provide a high-level representation of properties in multimodality but also comprehensively model the long-term and short-term temporal variability of underlying depression cues for precise depression estimation.

4.2 Attentional ConvBiLSTM

Generally, the model of Attention ConvBiLSTM resembles the model of Sub-Attentional ConvBiLSTM. Its architecture is illustrated in the Figure 4.2. As one can observe, it has the equivalent dimensionality of multimodality as model input, i.e. visual, audio, and text data, as well as the identical backbone for each feature branch. Moreover, 8 classification heads for the purpose of predicting PHQ-8 Subscores are also stacked at the end of the model. However, in contrast

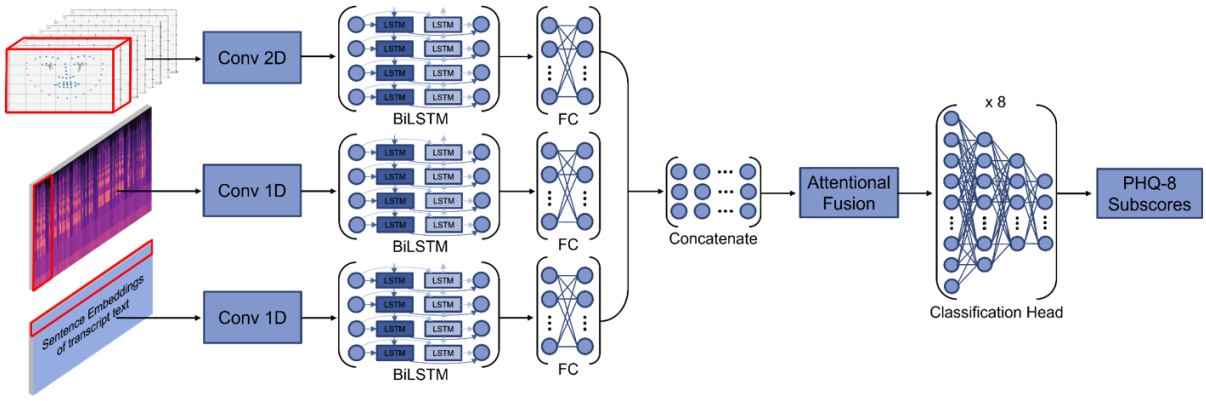


Figure 4.2: An architecture of Attentional ConvBiLSTM.

to Sub-Attentional ConvBiLSTM, the major difference is the number of attentional fusion layers. Here only one single attention fusion layer has been utilized, for the purpose of global attention of the concatenated feature map for each subclass, rather than local attention of each subclass with 8 attention fusion layers. The reason behind this model design is to reduce the complexity of the architecture while still proving to have a competitive performance compared with Sub-Attentional ConvBiLSTM. That is why this model is called Attentional ConvBiLSTM. Indeed, as demonstrated through the experiments, the Attentional ConvBiLSTM model yields the second-best performance throughout this thesis. For the details of the structure, functionality, and parameter settings in the identical network parts, please take a look at the description in the Section 4.1.

4.3 Attentional Fusion Layer

So far the overall architecture of the models has been discussed in-depth. This section will be focused on discussing the structure and functionality of the attentional fusion layer, exploited in the late fusion in the model. An illustration of its architecture is shown in the Figure 4.3. This attentional fusion layer is based on the work from Dai et al. [8], which has a hierarchical structure from top to bottom, splitting into 3 different layers.

In the first layer, given a feature map concatenated from extracted features of each modality $\mathbf{Y} \in \mathbb{R}^{C \times H \times W}$, it will first be processed by a 2-dimensional convolutional neural network, which will learn to capture and detect the most relevant and critical features to form a new local translation-equivariant response with the identical size of $(C \times H \times W)$. This response will then be added together with the input feature map $\mathbf{Y}$ to form the intermediate feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ as an input for the attentional block in the second layer. C at here denotes channels, which in our case is 1, and $H \times C$ denotes the size of the feature map.

In the second layer, given an intermediate feature map generated from the first layer, the output channel attention weight $w \in \mathbb{R}^{C}$ will be computed as

$$w = \sigma(G(\mathbf{X}) \oplus L(\mathbf{X}))), \tag{4.1}$$

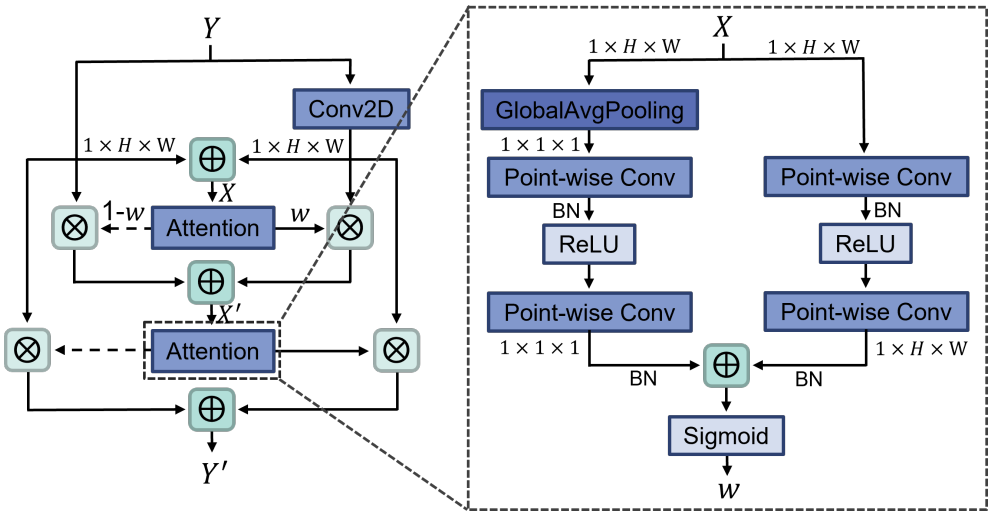


Figure 4.3: An architecture of the attentional fusion layer.

which is an aggregation of the global feature attention $G(\mathbf{X}) \in \mathbb{R}^C$ and the local channel attention $L(\mathbf{X}) \in \mathbb{R}^{C \times H \times W}$ transformed through sigmoid activation function σ as demonstrated in the left side and right side of the magnified image on the right in the Figure 4.3, according to the author Dai et al. [8]. The global feature attention can be obtained with the following equation

$$G(\mathbf{X}) = BN(W_2 ReLU(BN(W_1 g(\mathbf{X})))) \tag{4.2}$$

As the name of this functional component implies, the global feature context of the intermediate feature map will be first extracted through a global average pooling block (GAP) $g(\mathbf{X}) = \frac{1}{H \times W} \Sigma_{i=1}^{H} \Sigma_{i=1}^{W} \mathbf{X}_{[:,i,j]}$, followed by a dimension decreasing block $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, a rectified linear unit $ReLU$, and finally a dimension increasing block $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. r is the channel reduction ratio and both dimension decreasing and increasing block are in fact a point-wise convolution ($PWConv$). After each block, batch normalization BN is applied to regularize the intermediate representation of the extracted features to a standard normal distribution. As for the local feature attention, similar structure can be perceived excluding the global average pooling block (GAP) $g(\mathbf{X})$. Hence, the function could be summarize to

$$L(\mathbf{X}) = BN(PWConv_2 ReLU(BN(PWConv_1(\mathbf{X})))) \tag{4.3}$$

For each point-wise convolution, the kernel size, as well as the stride size, are both set to 1 and without any padding. Moreover, It is noteworthy that the outcome local attentional weight of L(X) has the identical shape as the input, which can be trained to preserve and highlight the subtle details of depression cues from the intermediate feature map. After the channel attention weight, w is derived, the complementary channel attention weight $(1 - w)$ is also calculated denoted with the dashed line in Figure 4.3. These two weights will then be multiplied with the input feature map $\mathbf{Y}$ and the translation-equivariant response extracted from the 2-dimensional CNN block $Conv(\mathbf{Y})$, respectively, to generate the refine feature (RF) as well as the complementary refine feature (RF^c), which could be expressed as follows:

$$\begin{aligned} RF &= Conv(\mathbf{Y}) \otimes w = Conv(\mathbf{Y}) \otimes \sigma(G(\mathbf{X}) \oplus L(\mathbf{X}))) \\ RF^c &= \mathbf{Y} \otimes (1 - w) = \mathbf{Y} \otimes (1 - \sigma(G(\mathbf{X}) \oplus L(\mathbf{X}))) \end{aligned} \tag{4.4}$$

Finally, the output of the second layer $\mathbf{X}' \in \mathbb{R}^{C \times H \times W}$, which is a transitional attentional feature, can be obtained as the summation of both refined features.

$$\mathbf{X}' = Conv(\mathbf{Y}) \otimes w \oplus \mathbf{Y} \otimes (1 - w) \tag{4.5}$$

Last but not least, in the third layer, the attentional process explained in the second layer will be performed again to further improve as well as harvest the superior focus from the transitional attentional feature $\mathbf{X}'$ of multimodality for accentuating depressive characteristics. Therefore, the ultimate attentional feature fusion output $\mathbf{Y}' \in \mathbb{R}^{C \times H \times W}$ can be expressed as

$$\mathbf{Y}' = Conv(\mathbf{Y}) \otimes w' \oplus \mathbf{Y} \otimes (1 - w') \tag{4.6}$$

where w' is the channel attention weight outputted from $\mathbf{X}'$. In our model, the attentional feature fusion output $\mathbf{Y}'$ will then be input to the 8 classification head for the PHQ-8 Subscores classification prediction.

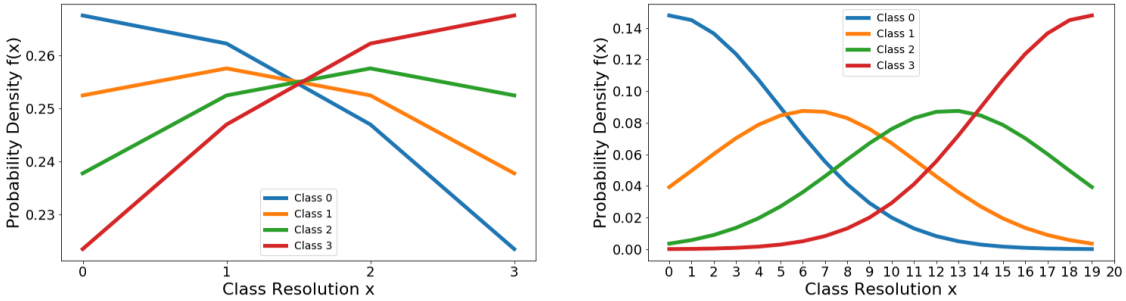
4.4 Multi-path Uncertainty-aware Score Distributions Learning (MUSDL)

In this work, rather than using the hard-label GT with the Cross-Entropy loss, the soft-label score distribution of GT, transformed with MUSDL, is exploited alongside the KL divergence loss. This method is inspired by the work from Tang et al. [35] and has demonstrated considerable effectiveness to solve the intrinsic ambiguity in GT scores and boost the model performance. Multi-path in the name indicates the strategy to explore the disentangled components of a GT score such as our PHQ-8 Subscores, which consists of 8 different subclasses, corresponding to 8 unique major depression symptoms, with each subclass having a class resolution of 4 ranging from 0 to 3.

Therefore, given a classification ground truth $\mathbf{s} \in \mathbb{N}^n$ of an interview clip contained a set of n hard-label scores $s \in \mathbb{N}$: $\mathbf{s} = [s_1, s_2, \ldots, s_n]$, each score s in the ground truth will be transformed into a gaussian distribution-like soft-label $\mathbf{s'} \in \mathbb{R}^{m'}$ by generating a Gaussian function with the mean of each s itself and a standard deviation of σ as follows:

$$\mathbf{s'} = g(s') \mid_s = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(s' - s)^2}{2\sigma^2}), \tag{4.7}$$

where the hard-label $\{s \in \mathbb{N} \mid 0 \leq s < m\}$ is an integer and the soft-label $\mathbf{s'} = [s'_1, s'_2, \ldots, s'_{m'}]$ a discrete vector with $\{s' \in \mathbb{R} \mid 0 \leq s' \leq 1\}$. Here σ is a hyper-parameter which serves as the level of uncertainty for assessing a clip and $m, m' \in \mathbb{N}$ denote the class resolution or number of the class before and after the soft-label transformation. The transformed ratio $r \in \mathbb{R}$ can



(a) Soft-label with $r = 1$. **(b)** Soft-label with $r = 5$.

Figure 4.4: Comparison of soft-label with different ratio.

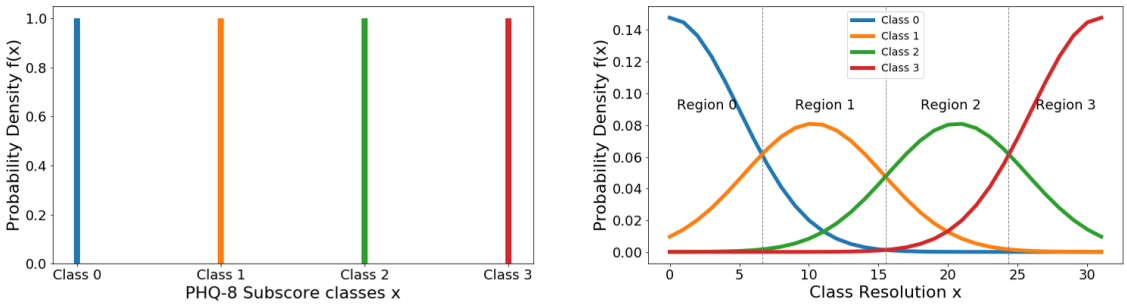
be derived with $r = \frac{m'}{m}$, which should be equal or greater than 1, indicating an unchanging or expansion of class resolution. The higher the ratio are, the smoother the distribution curve becomes, leading to a better soft-decision strategy performance. An illustration is shown in the Figure 4.4. The final soft-label distribution $\mathbf{s^{norm}} = [s_1^{norm}, s_2^{norm}, \ldots, s_{m'}^{norm}]$ of this particular hard-label value is generated by normalizing $\mathbf{s'}$ as below:

$$\mathbf{s^{norm}} = \mathbf{s'} / \sum_{i=1}^{m'} \mathbf{s'_i}, i = 1, 2, \ldots, m' \tag{4.8}$$

In the end, by uniformly discretizing each hard label s in the ground truth of this particular clip into a normalized soft-label vector $\mathbf{s^{norm}}$, a matrix of n gaussian distributions $\mathbf{S^{norm}} \in \mathbb{R}^{n \times m'}$ can be obtained. The overall transformation process can be summarized and expressed as the following equations:

$$\begin{aligned} \mathbf{s} &= [s_1, s_2, \ldots, s_n] \\ &\downarrow \text{ Label Transformation} \\ \mathbf{S^{norm}} &= [\mathbf{s_1^{norm}}, \mathbf{s_2^{norm}}, ..., \mathbf{s_n^{norm}}] \\ &= \begin{bmatrix} s_{1,1}^{norm} & s_{1,2}^{norm} & \cdots & s_{1,m'}^{norm} \\ s_{2,1}^{norm} & s_{2,2}^{norm} & \cdots & s_{2,m'}^{norm} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1}^{norm} & s_{n,2}^{norm} & \cdots & s_{n,m'}^{norm} \end{bmatrix} \end{aligned} \tag{4.9}$$

In this work, n is equal to 8 and m is 4 in accord with the definition of PHQ-8 Subscores. The standard deviation σ is set to 5 and the transformed ratio is 8, indicating that the number of the class is expended from $m = 4$ to $m' = 32$. The final transformed label is illustrated in the Figure 4.5. One can notice that before the transformation, the hard-label GT of 4 different classes is given. After the transformation, a probability density function of the normal distribution is generated. Furthermore, during the training stage, all of the 8 different classification head is trained to predict the probability between the 4 different depressive classes of the corresponding



(a) GT hard-label.

(b) Transformed GT soft-label.

Figure 4.5: An overview of GT transformation of MUSDL. Hard-label (a) und Soft-label (b).

subscore with the softmax-function: $\mathbf{S^{pred}} = [\mathbf{s_1^{pred}}, \mathbf{s_2^{pred}}, \ldots, \mathbf{s_n^{pred}}]$. The learning loss is then calculated through KL divergence between $\mathbf{S^{norm}}$ and $\mathbf{S^{pred}}$, which could be computed as:

$$KL(\mathbf{S^{norm}}, \mathbf{S^{pred}}) = \sum_{i=1}^{m'} \mathbf{s_i^{norm}} log \frac{\mathbf{S_i^{norm}}}{\mathbf{S_i^{pred}}} \tag{4.10}$$

As for the inferring phase, the predicted probability of each class under all PHQ-8 Subscores is derived from the well-trained model and the final assessment $\mathbf{s_{final}^{pred}}$ is obtained by selecting the score with the maximum probability in each subscore, then dividing by the ratio r and rounding down:

$$\mathbf{s_{final}^{pred}} = \lfloor \arg\max_{s_i} \{\mathbf{s_1^{pred}}, \mathbf{s_2^{pred}}, \ldots, \mathbf{s_n^{pred}}\}/r \rfloor \tag{4.11}$$

4.5 Sharpness-Aware Minimization (SAM)

As mentioned previously, one of the potential problems that this thesis is facing is the small-scale dataset, which hinders the model to have the generalization capability. Therefore, motivated by Chen et al. [7], the Sharpness-Aware Minimization (SAM) designed by Foret et al. [15], a second-order optimization technique, is executed in this work to improve the generalization of our model so that it could have a representative performance even just been trained on a small dataset. The problem with the first-order optimization is that even though it minimizes the training loss L_{train}, it dismisses the higher-order information such as curvature which correlates with the generalization, leading to a higher generalization error in text loss L_{test} according to [7]. Therefore, SAM is devised to take such a problem into consideration.

Intuitively, SAM seeks to find the weight parameter w of a model whose entire neighbors in the range ρ have low training loss L_{train} compared with other weight parameters, as stated by Chen et al. [7]. This interpretation could be formulated into a minimax decision shown below:

$$\min_{w} \max_{\|\varepsilon\|_2 \leq \rho} L_{train}(w + \varepsilon), \tag{4.12}$$

which is a second-order problem. However, due to the complexity of solving the exact inner maximization with the optimum ε_{opt}, Foret et al. [15] employ the first-order approximation for better efficiency of calculating the sharpness aware gradient $\hat{\varepsilon}(w)$, which could be structured as:

$$\begin{aligned}\hat{\varepsilon}(w) &= \underset{\|\varepsilon\|_2 \leq \rho}{\arg\max}\, L_{train}(w) + \varepsilon^T \nabla_w L_{train}(w) \\ &= \rho \nabla_w L_{train}(w) / \parallel \nabla_w L_{train}(w) \parallel_2,\end{aligned} \tag{4.13}$$

After the $\hat{\varepsilon}(w)$ is derived, SAM updates the current weight w based on the $\hat{\varepsilon}(w)$ with the following equation:

$$w' = \nabla_w L_{train}(w) \mid_{w+\hat{\varepsilon}(w)} \tag{4.14}$$

Figure 4.6 illustrates an example of the loss landscape for the same training process with and without SAM according to Foret et al. [15]. One can notice that for the one without SAM in the Figure 4.6(a), there is plenty of sharpness as well as local minima, causing the training weight to be trapped inside and thus losing the ability of generalization and the possibility of reaching the global minimum. The Figure 4.6(b), on the other hand, demonstrates a significant improvement in smoothing out these local minima in the loss landscape by implementing SAM. Therefore, it can be concluded that with this technique, the model can possibly be trained to reach closer to the global minimum, implying a greater model representation.

(a) Loss landscape without SAM. **(b)** Loss landscape with SAM.

Figure 4.6: Comparison of loss landscape before and after applying SAM according to [15].

Chapter 5

Experiments

In this chapter, all of the results from the experiments will be presented in detail, as well as how our best model and second-best model are derived step by step through several different comparison studies. Starting with Section 5.1, the overall experimental methodology and strategy will be introduced. In the Section 5.2, the unsuccessful attempts, where the model is trained on the generated raw dataset mentioned in the Section 3.5, will be discussed along with the explanation. In the Section 5.3, the results of all the single-modal models trained on the generated clipped dataset will be shown. Finally, the chief part and the central idea of this work, the multi-modal model, will be explored in the Section 5.4, followed by a chapter summary of the final model for depression estimation in Section 5.5.

5.1 Experimental Methodology

The overall experimental methodology in this thesis is to first train each single-modal model, for the purpose of training an effective backbone as a feature extractor for each modality as well as a branch in the multi-modal model and applying transfer learning, since the multi-modal model itself is huge and complex to train its model weights from scratch. After exploiting the pre-trained weights from the single-modal models to the multi-modal model, only the attentional fusion layer, as well as the 8 classification heads, are trained in the first learning phase. In the second phase, the whole multi-modal model is then fine-tuned together to further improve the model performance.

In addition to that, it is started by utilizing the generated raw dataset, which includes the whole interview data of each participant, to train the model for depression estimation. With this method, the model imitates the way how the interviewers perceive and diagnose major depression. However, while training every single modality, it is soon realized that this attempt fails in most of the experiments due to the error, noises, imbalanced data, etc. These all unsuccessful results are included in the Section 5.2. Therefore, our focus is then shifted to the second generated dataset, namely the clipped dataset mentioned in the Section 3.5. This time, it shows significant success in all experiments, and thus for the following Sections 5.3 and 5.4, the outcomes are all based on training on the clipped dataset. Moreover, for the audio single-modal model and the audio branch in the multi-modal model, the specifically generated audio dataset is employed to obtain a non-bias audio feature extractor, which indeed demonstrates a considerable model performance.

Last but not least, the most effective and low-memory cost backbone, advantageous decision strategy (hard-decision with hard-label or soft decision with soft-label), better fusion method,

etc. are also researched in this work. Hence, the overall experimental variables in each model part are listed in the Figure 5.1.

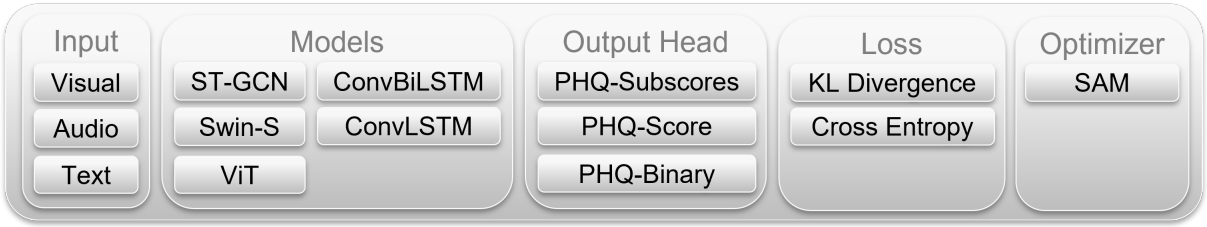


Figure 5.1: Overall experimental variable.

5.2 Unsuccessful Attempts

In this section, the depression estimation models are trained on the generated raw audio dataset, meaning that the data of the whole interview length in each modality will be input to train the model. Moreover, inspired by the current popular general-purpose CV backbone model such as Vision Transformer (ViT) [14], Swin Transformer [29], and Spatial-Temporal Graph Convolutional Networks (ST-GCN) [39], these backbone are utilized at here for different modalities. Starting with text modality, Vision Transformer (ViT) [14], is exploited to process the sentence embeddings by considering them as a gray-scale feature map. For audio modality, Swin Transformer devised by Liu et al. [29], is applied to extract features from the log-mel spectrogram. As for the visual modality, ST-GCN [29], is trained to generate the graph and harvest the underlying depression characteristics from micro-facial expression.

However, after several attempts, all experiments fail no matter how the parameters and variables are set due to the potential problems mentioned in the Section 3.3 and the fact that no techniques have been applied to solve and handle these issues in this generated raw dataset. Therefore, it is conspicuous and anticipatable that all models have a really strong bias toward non-depressed participants and only derive class 0 (without suffering from MD) as the prediction on the test dataset. Furthermore, during the experiments, some other problems emerge. One obvious problem that can directly be realized is the huge computational load because of the enormous data size, leading to several difficulties, e.g. inefficient training process, infinite time to train, running out of memory problems, etc. The interview length is ranging from 7 minutes to 33 minutes with an average of 16 minutes. This indicates that if the whole interview is needed to load, for instance, visual data of micro-facial expression, it will cost at least $30\,\mathrm{GB}$ to train the model, which is unrealistic. Therefore, this dataset is soon being abandoned and the focus of this thesis is shifted to the clipped dataset, which is predominantly utilized in the following Section 5.3 and Section 5.4.

5.3 Single Modality

As mentioned previously, in order to transfer the well-trained learning weights of the most effective feature extractor for different modalities to each backbone in the multi-modal models, different single-modal models with various set-ups have first been created, trained, tested, and compared based on the clipped dataset. Therefore, in this section, all the results as well as performances from these single-modal models with be discussed. The main score used to compare the classification metrics between each other is the accuracy of having MD.

Furthermore, throughout this stage, the answers to the following questions would like to be researched. The first question is "which prediction score is better?" Even though it has been stated in the Section 3.1 that the PHQ-8 Subscores are predominantly exploited in this thesis according to the definition of the PHQ-8 system, some comparison studies are still conducted to prove the correctness of this idea. The second question is "which input data works better?" For each modality, data has been transformed into different data types or data formats to test and ascertain the most informative input data type for depression estimation. The third question is "which is the most effective backbone as a feature extractor?" Inspired by several different models, including NLP and CV approaches, different backbones have been trained and tested to seek the most effective and low-memory cost feature extractor. Finally, the last question is "which decision strategy is better? soft-decision or hard-decision." To determine that, a comparison study between soft-decision with KL divergence loss and hard-decision with cross-entropy loss is also performed.

5.3.1 Text Modality

Starting with text modality, the sentence embeddings have been trained with several backbones such as ConvBiLSTM, Swim Transformer [29], ViT [14], etc., and different output heads corresponding to different predicted score in PHQ-8 system for depression estimation. However, most of the models fail by either showing a lack of convergence during the training phase or performing poorly on the test dataset with random guessing no matter what kind of backbone or output head is applied. Hence, only those models that train with ConvBiLSTM and predict the PHQ-8 Subscores are shown in the Table 5.1. One can notice that both models achieve relatively low accuracy regardless of the application of the soft-decision technique and the best accuracy is only 50.82%, which is merely higher than binary guessing. Therefore, summing up all the results shown and unshown here, it can be concluded that the text dataset itself provides insufficient information related to depression estimation.

Table 5.1: Text analysis

Modality	Data Type	Backbone	Output Head	Soft-Decision	Accuracy %
Which decision strategy is better?					
Text	SentEmbed (dim 512)	Conv1D-BiLSTM	PHQ-8 Subscores	No	50.82
Text	SentEmbed (dim 512)	Conv1D-BiLSTM	PHQ-8 Subscores	Yes	46.33

5.3.2 Visual Modality

Table 5.2 summarizes all the experiment results of visual modality.

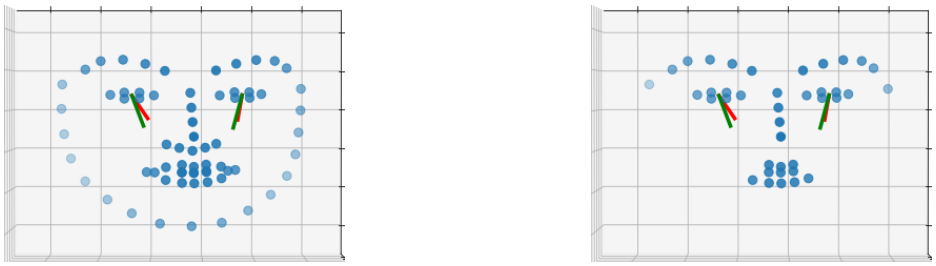
Table 5.2: Visual analysis

Modality	Data Type	Backbone	Output Head	Soft-Decision	Accuracy %
Which prediction score is better?					
Visual	KPs + Gaze	Conv2D-BiLSTM	PHQ-8 Subscores	No	76.94
Visual	KPs + Gaze	Conv2D-BiLSTM	PHQ-8 Score	No	51.22
Visual	KPs + Gaze	Conv2D-BiLSTM	PHQ-8 Binary	No	71.63
Which input data works better?					
Visual	KPs + Gaze	ST-GCN	PHQ-8 Subscores	No	55.71
Visual	KPs-39 + Gaze	ST-GCN	PHQ-8 Subscores	No	43.67
Which decision strategy is better?					
Visual	KPs + Gaze	Conv2D-BiLSTM	PHQ-8 Subscores	No	76.94
Visual	KPs + Gaze	Conv2D-BiLSTM	PHQ-8 Subscores	Yes	**79.59**

In the first part of the Table 5.2, the ConvBiLSTM models are trained with cross-entropy loss and seek to find out which prediction score is the best. By comparing the outcome of those three, it is proved that utilizing an output head for predicting PHQ-8 Subscores yields the best performance with an accuracy of around 77% in comparison with the other two, namely PHQ-8 Score and PHQ-8 Binary. Therefore, for the following experiments in visual single-modal models, only PHQ-8 Subscores are utilized as output heads.

Moreover, inspired by the current state-of-the-art CV technique in handling key points data, ST-GCN [39] model is also employed as a backbone to make use of graph convolutions known from the spectral graph theory [20]. But one can notice that the performance is not uncomparable to the ConvBiLSTM model with solely 55.71% accuracy. Hence, to solve this and further improve the model, a reduction of facial key points number is carried out since the original ST-GCN model is devised to construct the graph for less than 25 key points. In this work, however, there are 68 facial key points plus 4 gaze direction vectors included in the visual dataset, meaning a total of 72 key points, which might be too complicated for the model as its scale is not that large enough to process such complexity. The number of key points is deducted until 43 with 39 facial key points and 4 gaze direction vectors by removing the less important key points, which shows less change in motion. An illustration is demonstrated in the Figure 5.2. Mainly, the contour of each face, as well as the under-lip, are erased. Nonetheless, as opposed to the expectation, the model even yields a lower accuracy than the previous one without key points reduction. Hence, it is convinced that the ConvBiLSTM is the most effective backbone for extracting micro-facial expression features.

Finally, to further improve the visual single-modal model, MUSDL is executed. By converting the GT to soft-label and computing the loss with KL divergence for soft-decision, an around 3%



(a) Before. **(b)** After.

Figure 5.2: An illustration of facial key points reduction. The Figure 5.2(a) is visual data with 72 key points and The Figure 5.2(b) is 43 key points.

model performance improvement can be perceived in the last part of the Table 5.2, indicating the importance of this technique for achieving a better depression estimation model.

To sum up, the best model in visual modality single-modal model is the ConvBiLSTM model trained with soft-decision and an output head predicting PHQ-8 Subscores, which accomplishes an accuracy of 79.59% for MD detection marked in bold in the Table 5.2.

5.3.3 Audio Modality

The last single modality to train is audio data. Here, different output heads, input data types, backbone feature extractors, and the application of soft-decision are also going to be trained and tested based on the generated audio dataset, which is applied with the gender balancing technique. Furthermore, to discover the effectiveness of such a technique, the ConvBiLSTM model trained on the original audio dataset is also implemented to conduct a comparison study. The summary of all the experiments is organized in the Table 5.3.

Similar to the experiment in the visual modality, the different predicted scores in the PHQ-8 system for MD detection are first trained and compared with the ConvBiLSTM backbone to prove the assumption that the model performs the best with the PHQ-8 Subscores. In fact, as demonstrated in the first part of the Table 5.3, the output head for predicting PHQ-8 Subscores also achieves a higher accuracy of over 10% in comparison with the others. Moreover, the log-spectrogram as the input data type is also trained with the ConvBiLSTM model to find out the answer of which input data type works better. Previously, according to the issues observed in the visualization of log-spectrograms and potential problems mentioned in the Subsection 3.4.3, it is assumed that the log-mel spectrogram is better. Now based on the experimental results by comparing the first and second part in Table 5.3, it is to confirm that the log-mel spectrogram, indeed, provides more informative data for the model, leading to a higher model accuracy generally.

For pure CV technique, this time the current popular model, Swin Transformer [29], which serves capably severs as the general-purpose backbone for computer vision, is exploited, particularly with the small-scale Swin Transformer (Swin-S). It is trained with the three different output heads and cross-entropy loss function for hard-decision. However, the results are not promising enough with relatively low accuracy. The model with PHQ-8 Binary output head even fails by

Table 5.3: Audio analysis

Modality	Data Type	Backbone	Output Head	Soft-Decision	Accuracy %
Which prediction score is better?					
Audio	Mel Spectrogram	Conv1D-BiLSTM	PHQ-8 Subscores	No	76.53
Audio	Mel Spectrogram	Conv1D-BiLSTM	PHQ-8 Score	No	64.69
Audio	Mel Spectrogram	Conv1D-BiLSTM	PHQ-8 Binary	No	63.67
Which input data works better?					
Audio	Spectrogram	Conv1D-BiLSTM	PHQ-8 Subscores	No	69.80
Audio	Spectrogram	Conv1D-BiLSTM	PHQ-8 Score	No	45.51
Audio	Spectrogram	Conv1D-BiLSTM	PHQ-8 Binary	No	45.31
Which is the most effective model as feature extractor?					
Audio	Mel Spectrogram	Swin-S	PHQ-8 Subscores	No	51.84
Audio	Mel Spectrogram	Swin-S	PHQ-8 Score	No	31.02
Audio	Mel Spectrogram	Swin-S	PHQ-8 Binary	No	failed
Which decision strategy is better? + Impact of audio gender balancing (GB)					
Audio	Mel Spectrogram + No GB	Conv1D-BiLSTM	PHQ-8 Subscores	No	70.00
Audio	Mel Spectrogram	Conv1D-BiLSTM	PHQ-8 Subscores	No	76.53
Audio	Mel Spectrogram	Conv1D-BiLSTM	PHQ-8 Subscores	Yes	**76.73**

predicting only class 0 as output for each clip, meaning that the model does not learn and predict at all. The reason for this is that unlike normal images or videos, which have certain and relative apparent objects, the log-mel spectrogram consists of solely signal of the acoustic spectrum as illustrated in the Figure 3.9, whose spatial as well as the temporal difference with each other is genuinely subtler then normal images and videos. These subtle differences are too micro and thus are too complex for Swin Transformer, which is solely designed for general CV tasks, whereas, in depression estimation, it is much more like an NLP task, considering the interview as a time series data. Therefore, this approach is then soon put aside and the attention is shifted back to the ConvBiLSTM model.

Last but not least, to understand how the gender balancing technique benefits the model performance, a comparison experiment is implemented and the outcome is shown in the fourth part of the Table 5.3. It is noticeable that the ConvBiLSTM model performance gets boosted by over 6.5% by applying the gender balancing technique. Before GB, the model only has an accuracy of 70%. But after GB, it reaches around 76.5% accuracy. This implies that the acoustic gender bias problem really exists and it has been tackled in this thesis through the data preprocessing. To further improve the performance, soft-decision is exploited and a slight advancement can be observed in the last line.

In summary, resembling the visual single-modal model, the best backbone for audio modality single-modal model is the ConvBiLSTM model, followed by an output head predicting PHQ-8 Subscores and training with soft-decision strategy. The audio single-modal model achieves an accuracy of 76.73% for MD detection marked in bold in the Table 5.3.

5.4 Multimodality

So far all the single-modal models have been completed and each well-trained learning weights of the backbone from each modality will be transferred to the multi-modal model accordingly. Moreover, it has also been proven that the ConvBiLSTM model backbone is the most effective feature extractor regardless of which modality and PHQ-8 Subscores is the better prediction score for depression estimation. Therefore, these variable settings are continued to be utilized, indicating that for each multi-modal experiment in this section, only the ConvBiLSTM model with 8 classification output heads for predicting PHQ-8 Subscores is exploited. This section aims to focus on seeking out the most effective fusion methods as well as the impact of the soft-decision strategy.

In total, 8 different fusion methods have been tested in this work, which could be categorized into the traditional fusion method and the weighting fusion method. In terms of the key difference between these two, the weighting fusion method differs from the traditional fusion method in training an extra learning layer for attention. For the traditional fusion method, the following six feature fusion approaches are used:

- Concatenation method: $\mathbf{y_{nd}^{cat}} = f^{cat}(\mathbf{x_d^1}, \mathbf{x_d^2}, \cdots, \mathbf{x_d^n})$
- Summation method: $\mathbf{y_d^{sum}} = \mathbf{x_d^1} + \mathbf{x_d^2} + \cdots + \mathbf{x_d^n}$
- Mean method: $\mathbf{y_d^{mean}} = (\mathbf{x_d^1} + \mathbf{x_d^2} + \cdots + \mathbf{x_d^n})/n$
- Median method: $\mathbf{y_d^{median}} = median(\mathbf{x_d^1}, \mathbf{x_d^2}, \cdots, \mathbf{x_d^n})$
- Maximum method: $\mathbf{y_d^{max}} = max(\mathbf{x_d^1}, \mathbf{x_d^2}, \cdots, \mathbf{x_d^n})$
- Multiplication method: $\mathbf{y_d^{multi}} = \mathbf{x_d^1} \otimes \mathbf{x_d^2} \otimes \cdots \otimes \mathbf{x_d^n}$

, where n denoted the number of extracted vectors and d indicates the dimension of each extracted vector. Each method besides concatenation is an element-wise operation. As for the weighting fusion method, the following two feature fusion approaches are employed:

- Sub-Attentional fusion method
- Attentional fusion method

5.4.1 AV Modality

Since the visual and audio single-modal models yield the greatest success in single modality based on the ConvBiLSTM model with the achieved accuracies of 79.59% and 76.73%, respectively, the first idea that comes to the mind is to fuse these two modalities. Hence, the Audio+Visual (AV) multi-modal model is built and tested. The final results are summarized in the Table 5.4.

Overall, it can be noticed that the multi-modal feature fusion increases the model performance in comparison with the single-modal feature. Before fusing multi-modal features, the highest

Table 5.4: Audio+Visual (AV) analysis

Modality	Fusionmethod	soft-decision	Accuracy %
Which fusion method is most effective?			
Audio + Visual	Concatenate	No	79.59
Audio + Visual	Concatenate	Yes	79.80
Audio + Visual	Sum	No	80.82
Audio + Visual	Sum	Yes	81.43
Audio + Visual	Mean	No	81.22
Audio + Visual	Mean	Yes	81.02
Audio + Visual	Median	No	79.80
Audio + Visual	Median	Yes	80.20
Audio + Visual	Max	No	80.41
Audio + Visual	Max	Yes	80.82
Audio + Visual	Multi	No	79.80
Audio + Visual	Multi	Yes	79.80
Audio + Visual	Attention	No	**82.04**
Audio + Visual	Attention	Yes	81.43
Audio + Visual	Sub-Attention	No	81.63
Audio + Visual	Sub-Attention	Yes	**82.04**

accuracy that can be achieved is only 79.59% in the visual modality. Now, most of the performances of multi-modal models research over 80% or even 81% accuracy, which is fairly high in depression estimation tasks considering how many patients with depression are not recognized in the current clinical statistics [22, 32].

Furthermore, it is worth mentioning that the weighting fusion methods generally perform better than the traditional fusion methods with a 1% accuracy improvement on average according to the Table 5.4. For each fusion method, a soft-decision strategy is also employed, which leads to a slight advancement in most of the models. The best traditional fusion method, in this case, is the summation method with an accuracy of 80.82% without applying MUSDL and 81.43% with applying MUSDL and the best weighting fusion method can not be known for certain as they both show comparable outcomes and yield the highest accuracy, 82.04%, in AV modality. However, by comparing the other evaluation metric scores, which are not shown here, the Attentional ConvBiLSTM model is slightly better than the Sub-attentional ConvBiLSTM model in AV modality.

To sum up, the best multi-modal model in AV modality is the Attentional ConvBiLSTM model, which achieves the highest accuracy of 82.04% marked bold in the Table 5.4. It utilizes ConvBiLSTM as the backbone for each input modality to effectively extract the features and fuse these features with a general attentional fusion layer, followed by 8 different classification output heads for the purpose of predicting PHQ-8 Subscores.

5.4.2 AVT Modality

Even though it has been shown in the text single-modal model that the text data itself provides inadequate information related to depression estimation, it is to wonder that it may still contribute some when it is evaluated and fused with other modalities. Therefore, the Audio+Visual+Text (AVT) analysis is conducted and the outcomes are listed in the Table 5.5.

Table 5.5: Audio+Visual+Text (AVT) analysis

Modality	Fusion method	soft-decision	Accuracy %
Which fusion method is most effective?			
Audio + Visual + Text	Concatenate	No	80.00
Audio + Visual + Text	Concatenate	Yes	80.82
Audio + Visual + Text	Sum	No	80.82
Audio + Visual + Text	Sum	Yes	81.22
Audio + Visual + Text	Mean	No	81.63
Audio + Visual + Text	Mean	Yes	80.00
Audio + Visual + Text	Median	No	81.63
Audio + Visual + Text	Median	Yes	82.04
Audio + Visual + Text	Max	No	80.61
Audio + Visual + Text	Max	Yes	81.22
Audio + Visual + Text	Multi	No	80.41
Audio + Visual + Text	Multi	Yes	80.00
Audio + Visual + Text	Attention	No	81.22
Audio + Visual + Text	Attention	Yes	82.25
Audio + Visual + Text	Sub-Attention	No	82.04
Audio + Visual + Text	Sub-Attention	Yes	**82.65**

By fusing the text features, it is noticed that the overall performance raises around 0.5 - 1% no matter which fusion technique is used. Moreover, the weighting fusion methods still show higher performance compared with traditional fusion methods, indicating that an extra training layer for attentional feature fusion does provide advantages in harvesting deeper underlying depression clues for a better depression estimation. This time, however, the best traditional fusion method is the median method with an accuracy of 81.63% without applying MUSDL and 82.04% with applying MUSDL, which almost catches up with the performance of weighting fusion methods in AV modality. On the other hand, the best weighting fusion method, in this case, is the sub-attentional fusion approach with 82.04% accuracy without MUSDL and 82.65% accuracy with MUSDL.

In conclusion, the best AVT multi-modal model is the Sub-attentional ConvBiLSTM model, which yields the highest accuracy of 82.65% by fusing three different modalities, i.e. visual data of micro-facial expression, audio data of log-mel spectrogram, and text data of sentence embeddings, and trained with KL divergence for soft-decision loss. This result is marked bold in the Table 5.5. The Sub-attentional ConvBiLSTM model consists of three ConvBiLSTM

backbones for each modality and 8 different attentional fusion layers for subclass attention connected with 8 classification output heads, respectively, to predict PHQ-8 Subscores.

5.5 Summary of Final Model

So far plenty of different models, as well as experiments, have been discussed in-depth, including single modality and multimodality. These might cause some confusion as well as obscure the final chosen model in each part. Therefore, in this section, a brief summary is made as listed below for the reader to have a better overview of the final models, which are strongly related to the discussions in the following Chapter 6.

- Audio single-modal model: ConvBiLSTM model with 76.73% accuracy
- Visual single-modal model: ConvBiLSTM model with 79.59% accuracy
- AV multi-modal model: Attentional ConvBiLSTM model with 82.04% accuracy
- AVT multi-modal model: Sub-attentional ConvBiLSTM model with 82.65% accuracy

Chapter 6

Results and Discussions

To further understand how our models perform as well as estimate the depression severity between both genders and participants, two ablation studies have been carried out, i.e. "gender analysis" in the Section 6.1 and "participant analysis" in the Section 6.2. Moreover, a comparison between our approaches and the state-of-the-art methods is also summarized in the Section 6.3

6.1 Sensitivity of Gender Depression Estimation

The purpose of the gender analysis in this section is to dive deep into each modal and comprehend how sensitive each model is in terms of detecting MD between each gender and how significant the gender balancing technique is to suppress the gender bias phenomenon. Therefore, the predicted test results of all clips in the test dataset are categorized into female and male groups based on the gender GT of each clip. The female and male accuracies are then derived accordingly. In the Table 6.1, all the results for the gender analysis are summarized. One can notice that the overall model accuracy, female accuracy, and male accuracy are included inside. Furthermore, in the last column, the absolute value of the difference between female and male accuracies is also calculated to have an intuitive interpretation of the improvement from gender bias. The best score in each column is marked in bold.

Table 6.1: Gender analysis

Model Name	Modality	Acc %	Female - Acc %	Male - Acc %	Difference %
ConvBiLSTM	A (No GB)	70.00	64.44	77.67	13.23
ConvBiLSTM	A	76.73	79.23	73.3	5.93
ConvBiLSTM	V	79.59	79.93	79.13	0.80
Attentional ConvBiLSTM	AV	82.04	80.63	**83.98**	3.35
Sub-Attentional ConvBiLSTM	AVT	**82.65**	**82.39**	82.04	**0.35**

Starting from the first model in the Table 6.1, ConvBiLSTM trained on audio modality and without GB, it is to be observed that there is a huge accuracy difference of around 13% between both genders, indicating a strong gender bias phenomenon. However, by applying the gender balancing technique shown in the second model, a reduction of gender bias of around 7.3% has been yielded and now the difference between these two groups is solely around 6%. This signifies the seriousness of the role that the gender bias problem in the acoustic features plays and how critical it is to handle it during the audio preprocessing stage. Visual features, on the

other hand, show no problem of gender bias with a gender accuracy difference of less than 1%, which is also understandable as one can imagine how challenging it is for a person to distinguish a participant's gender solely based on the 68 3D facial key points illustrated in the Figure 3.2, let alone for computer.

Furthermore, by going down to the multi-modal model, a decreased tendency of gender accuracy difference in comparison to the audio single-modal model can be discovered, implying that the more different modalities are fused, the lower this audio gender bias phenomenon shows up. This is due to the fact that by fusing more and more data, the proportion, where the audio features account for, is getting lower and lower. Finally, with the Sub-attentional ConvBiLSTM model trained on AVT modality, the lowest gender accuracy difference is achieved, namely only 0.35%, and thus it has the highest sensitivity in gender depression estimation over 82% accuracy in both genders.

6.2 Sensitivity of Participants Depression Estimation

So far the models have been trained and tested on the clipped dataset, suggesting that they predict the depression severity for each clip rather than each participant. However, this will raise an argument concerning whether the result of each model is representative enough. Therefore, to resolve such doubt, the participant analysis is conducted in this section and the algorithm is created to recombine the clips into each participant based on each ID.

The pseudocode of the algorithm is demonstrated in the Algorithm 2. Generally, the GT ID of every clip in each batch are stored while the model processes and predicts the score for every clip. These two groups of values are kept in two different array but in parallel, indicating that they are matched based on the indices. Therefore, after the predicting stage, these predicted scores of every clip can be recombined together to form the original interview by iterating through each participant ID and utilizing the indices extracted from the GT ID array based on the current ID. After the reconstruction of the interview of each participant, the final PHQ-8 Score is computed as the mean of all clips of each participant and a threshold of 0.5 is set for the final PHQ-8 Binary, meaning that if over 50% of the clips of the current participant are being classified as depressed by the multi-modal model, then it can be concluded that this participant are having MD, and vice versa.

The final recalculated results from each model are summarized in the Table 6.2, where the clipped data accuracy, which has been seen and exploited in the previous sections, and partici-

Table 6.2: Participant analysis

Model Name	Modality	Clipped data - Acc %	Participant - Acc %	Improvement %
ConvBiLSTM	A (No GB)	70.00	70.21	0.21
ConvBiLSTM	A	76.73	78.72	1.99
ConvBiLSTM	V	79.59	78.72	-0.87
Attentional ConvBiLSTM	AV	82.04	80.85	-1.19
Sub-Attentional ConvBiLSTM	AVT	**82.65**	**85.11**	**2.46**

pant accuracy, which is the recalculated result, are included. Moreover, in order to intuitively understand how the participant accuracy differs from the clipped data accuracy, the improvement score are computed in the last column by subtracting the participant accuracy with the clipped data accuracy. Overall, the tolerance between both accuracies is relative low, less than 2.5% according to the Table 6.2. Therefore, it comes to the conclusion that all of the models, as well as the technique of training on the clipped dataset for depression estimation, are valid and representative.

Furthermore, one can perceive that there is even a performance improvement of around 2.5% in our best model, Sub-attentional ConvBiLSTM. This is actually due to the fact that even a participant is depressed or having MD, he or she will not consistently expresses depressive symptoms throughout the whole interview, for instance, usually there is a small talk at the beginning of each interview to break the ice. During this time, most of the participants will not show any depression symptom and can communicate properly, indicating that this clip from this interview should be, in fact, classified as non-depressed, even though the participant is depressed. Therefore, through this algorithm, this non-error mistake can be corrected. In addition to that, the ambiguous clips that have a PHQ-8 Score of 9 or 10 which is around the threshold and thus can confuse the model can be also rectified with the help of this algorithm. An illustration of participant analysis along with final scores and GT are demonstrated in the Figure 6.1.

ID	PHQ-8 GT		PHQ-8 Prediction		clip_01	clip_02	clip_03	clip_04	clip_05	clip_06	clip_07	clip_08	clip_09
	Binary	Score	Binary	Score									
P_453	1	17	1	17	21	15	21	19	19	17	12	20	13
P_323	0	1	0	2	3	0	0	0	2	2	2	7	
P_332	1	18	1	13	9	11	11	11	11	11			
P_421	1	10	1	12	13	9	12	11	16	17	8	13	
P_334	0	5	0	8	9	2	9	7	9	10	9	10	13
P_408	0	0	0	5	8	11	2	0	11	2	1		

Figure 6.1: A visualization of participant analysis.

Under each clip, the PHQ-8 Subscore and PHQ-8 Binary are denoted with the value inside the block and the color of the block, respectively. The gray background indicates the class 1, suffering from MD, whereas the white background indicates the class 0, does not suffering from MD. For the normal case, the Sub-attentional model predicts the whole clips from the interview of the participant as class 1 or 0, which is shown in the first two participants in the Figure 6.1. For the rest of the specific situations, owing to the aboeve-mentioned problem and ambiguous clips, one can notice a mix of predicted class 0 and class 1 of clips in each interview. This mix, however, is rectified through the algorithm as one can observe from the final binary status of MD, which is in accord with the GT.

Algorithm 2: Recombining clips back to participant.

Data: $dataloader$

Result: Returning predicted phq-binary and phq-score based on each participant

Initialize ID_gt, phq_score_pred, phq_binary_pred
for $batch$ *in* $dataloader$ **do**
    $id\_gt \leftarrow$ store $batch['id']$ in sequence
    $probs$ = ***ModelProcessing***($batch$)
    $pred\_score$, $pred\_binary$ = ***ComputeScore***($probs$)
    $phq\_score\_pred \leftarrow$ store $pred\_score$
    $phq\_binary\_pred \leftarrow$ store $pred\_binary$
end
% id_count corresponds to number of clip of particular ID
id_list, $id\_count \leftarrow$ extracting list of unique ID and each amount from id_gt
$threshold = 0.5$
for id *in* id_list **do**
    $indices \leftarrow$ where $id\_gt == id$
    % Sorting out particular clips to form whole interview of current participant
    $id\_clips\_score\_pred \leftarrow$ extracting from phq_score_pred with $indices$
    $id\_clips\_binary\_pred \leftarrow$ extracting from phq_binary_pred with $indices$
    % Recalculating the values $participant\_score\_pred$ = ***Mean***($id\_clips\_score\_pred$)
    $participant\_binary\_pred$ = ***Filter***($id\_clips\_binary\_pred$, id_count, $threshold$)
end

6.3 Final Model Analysis with State-of-the-Art Methods

Finally, in order to compare with the state-of-the-art approaches, the following scores are further derived: f1-score, precision, recall, MAE, and RMSE. The summary of the comparison is shown in the Table 6.3. Here, the single-modal models and multi-modal models are both included, along with different analysis approaches, namely clipped data-based as well as participant-based marked with †. The best score in each column of our approaches are marked in bold.

As one can observe, with the participant-based Sub-attentional ConvBiLSTM model with AVT multimodality, the best f1-score have been achieved in our approaches, namely 0.70, which is the second best compared with the other state-of-the-art models. Similarly, this model also yields the overall highest precision score, which is almost 0.9, meaning that for a participant being classified as having MD, this prediction has a 90% possibility to be correct, which are fairly high. As for the MAE and RMSE, the best model is the clipped data-based Attentional ConvBiLSTM model with AV multimodality. it achieve a MAE of 4.92 and a RMSE of 5.86, which are both the second best scores in comparison to the baseline models.

In summary, our model achieves a highly comparable result with the current state-of-the-art methods, particularly shown in AVT Sub-attentional multi-modal modal and AV Attentional ConvBiLSTM multi-modal model.

Table 6.3: A comparison with state-of-the-art methods

■ Comparison of SOTA		PHQ-8 Binary			PHQ-8 Score	
Method	**Modality**	**F1-Score**	**Precision**	**Recall**	**MAE**	**RMSE**
Baseline						
Valstar et al. [37]	A	0.46	0.32	0.86	5.36	6.74
Ma et al. [30]	A	0.52	0.35	1.00	-	-
Valstar et al. [37]	V	0.50	0.60	0.43	5.88	7.13
Williamson et al. [38]	V	0.53	-	-	5.33	6.45
Valstar et al. [37]	AV	0.50	0.60	0.43	5.52	6.62
Alhanai et al. [3]	AT	0.77	0.71	0.83	5.10	6.37
Gong et al. [16]	AVT	0.70	-	-	2.77	3.54
Our Approach						
ConvBiLSTM	A	0.61	0.56	**0.66**	5.19	6.93
ConvBiLSTM	V	0.61	0.64	0.58	6.17	8.06
Atten ConvBiLSTM	AV	0.61	0.79	0.58	**4.92**	**5.86**
†Atten ConvBiLSTM	AV	0.61	0.78	0.50	5.06	6.06
Sub-Atten ConvBiLSTM	AVT	0.65	0.73	0.58	4.99	6.67
† Sub-Atten ConvBiLSTM	AVT	**0.70**	**0.89**	0.57	5.04	6.98

† Participant-based

Chapter 7

Conclusion

In this chapter, the conclusion of this thesis and the outlook for the future work will be discussed in detail in the Section 7.1 and the Section 7.2, separately.

7.1 Conclusion

In this thesis, a highly accurate model based on multi-modal deep learning-based approach for automatic depression estimation has been realized. The best model throughout the work is the Sub-attentional ConvBiLSTM model with AVT multimodality based on the participant-based analysis, which achieves an overall accuracy of 85.11%, a f1-score of 0.70, and a precision of 0.89. It has also been proven through several different experiments that in general, the multi-modal model yields the better performance in comparison to single-modal model, ranging from 0.5% to 6% model accuracy improvement in this thesis. Furthermore, the applicability of exploiting micro-facial expressions for depression detection is also demonstrated with visual single-modal model yields an accuracy of 79.59%. The gender bias phenomenon in acoustic features have also been pointed out and solved by applying the gender balancing technique, which leads to an 6.7% overall model performance improvement and a reduction of gender accuracy bias of 7.3%. Last but not least, the utilization of sliding window technique to preprocess DAIC-WOZ database and training the model based on the clipped dataset have also been justified since the tolerance between the clipped data accuracy and the participant accuracy is under 2.5%, which is relative low, indicating a representative result.

In conclusion, our multi-modal model is proved to have the competitive performance with state-of-the-art approaches for depression estimation using the knowledge from micro-facial expression, audio, and text modality while facing with imbalanced and small-scale dataset problems.

7.2 Future Work

In this thesis, the ConvBiLSTM is utilized as the model backbone for each modality. This backbone itself, however, is relative outdated for NLP task to process the time series data even though it seems to work well. The current most popular state-of-the-art approach for NLP that comes into the question is the so-called Transformer, which relies entirely on self-attention mechanism to compute representations of sequence-to-sequence tasks. Therefore, from the

model-driven perspective, one interesting outlook for the future work will be to replace the ConvBiLSTM model backbone with the Transformer model to see how it can further raise the model performance by harvesting the deeper underlying depression cues with attentional mechanism.

Moreover, essential feature engineering techniques have been implemented in this thesis such as converting raw audio signal to log-mel spectrogram, gender balancing, and normalization. However, to further improve the cleanness of the dataset and to avoid non-bias features, the more complicated and deeper feature engineering techniques can be employed, e.g. topic modelling for text modality, deriving statistical visual features , and extracting non-gender bias acoustic features. Beside these, it has also been proven in this work that the more various modalities are fused together, the higher performance the model can achieve. Hence, from the more data-driven aspect, either more intensive feature engineering techniques can be conduct to generate a cleaner depression dataset or collecting more and more diverse data related to the major depression symptoms for the model training in the future work.

Bibliography

[1] Male to female voice changer. LingoJam.

[2] Depression key facts. World Health Organization, Sep 2021.

[3] Al Hanai, Tuka, Ghassemi, Mohammad M, and Glass, James R. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720, 2018.

[4] Bailey, Andrew and Plumbley, Mark D. Gender bias in depression detection using audio features. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 596–600. IEEE, 2021.

[5] Baltrušaitis, Tadas, Robinson, Peter, and Morency, Louis-Philippe. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[6] Cer, Daniel, Yang, Yinfei, Kong, Sheng-yi, Hua, Nan, Limtiaco, Nicole, John, Rhomni St, Constant, Noah, Guajardo-Cespedes, Mario, Yuan, Steve, Tar, Chris, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[7] Chen, Xiangning, Hsieh, Cho-Jui, and Gong, Boqing. When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

[8] Dai, Yimian, Gieseke, Fabian, Oehmcke, Stefan, Wu, Yiquan, and Barnard, Kobus. Attentional feature fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3560–3569, 2021.

[9] DAIC-WOZ Database. Accessed Oct. 21, 2019.

[10] Dalal, Navneet and Triggs, Bill. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[11] Degottex, Gilles, Kane, John, Drugman, Thomas, Raitio, Tuomo, and Scherer, Stefan. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014.

[12] Dham, Shubham, Sharma, Anirudh, and Dhall, Abhinav. Depression scale recognition from audio, visual and text analysis. *arXiv preprint arXiv:1709.05865*, 2017.

[13] Diagnostic, American_Psychiatric_Association. Statistical manual of mental disorders, 1994.

[14] Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Foret, Pierre, Kleiner, Ariel, Mobahi, Hossein, and Neyshabur, Behnam. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[16] Gong, Yuan and Poellabauer, Christian. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 69–76, 2017.

[17] Google. Universal sentence encoder large v5. *TensorFlow Hub*. Accessed 2018 [Online], 2018.

[18] Gratch, Jonathan, Artstein, Ron, Lucas, Gale, Stratou, Giota, Scherer, Stefan, Nazarian, Angela, Wood, Rachel, Boberg, Jill, DeVault, David, Marsella, Stacy, et al. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, 2014.

[19] Haque, Albert, Guo, Michelle, Miner, Adam S, and Fei-Fei, Li. Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*, 2018.

[20] Henaff, Mikael, Bruna, Joan, and LeCun, Yann. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[21] Hershey, John R and Olsen, Peder A. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE, 2007.

[22] Jacobi, Frank, Wittchen, H-U, Hölting, Christoph, Höfler, Michael, Pfister, Hildegard, Müller, Norbert, and Lieb, Roselind. Prevalence, co-morbidity and correlates of mental disorders in the general population: results from the german health interview and examination survey (ghs). *Psychological medicine*, 34(4):597–611, 2004.

[23] Kroenke, Kurt and Spitzer, Robert L. The phq-9: a new depression diagnostic and severity measure, 2002.

[24] Kroenke, Kurt, Strine, Tara W, Spitzer, Robert L, Williams, Janet BW, Berry, Joyce T, and Mokdad, Ali H. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.

[25] Lam, Genevieve, Dongyan, Huang, and Lin, Weisi. Context-aware deep learning for multi-modal depression detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3946–3950. IEEE, 2019.

[26] Lin, Lin, Chen, Xuri, Shen, Ying, and Zhang, Lin. Towards automatic depression detection: A bilstm/1d cnn-based model. *Applied Sciences*, 10(23):8701, 2020.

[27] Littlewort, Gwen, Whitehill, Jacob, Wu, Tingfan, Fasel, Ian, Frank, Mark, Movellan, Javier, and Bartlett, Marian. The computer expression recognition toolbox (cert). In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 298–305. IEEE, 2011.

[28] Liu, Hong, Gao, Yuan, and Wu, Pinging. Smile detection in unconstrained scenarios using self-similarity of gradients features. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1455–1459. IEEE, 2014.

[29] Liu, Ze, Lin, Yutong, Cao, Yue, Hu, Han, Wei, Yixuan, Zhang, Zheng, Lin, Stephen, and Guo, Baining. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[30] Ma, Xingchen, Yang, Hongyu, Chen, Qiang, Huang, Di, and Wang, Yunhong. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 35–42, 2016.

[31] Saidi, Afef, Othman, Slim Ben, and Saoud, Slim Ben. Hybrid cnn-svm classifier for efficient depression detection system. In *2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, pages 229–234. IEEE, 2020.

[32] Schrift, AA and Notfall, Hilfsangebote im. Zahlen und fakten.

[33] Song, Siyang, Shen, Linlin, and Valstar, Michel. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 158–165. IEEE, 2018.

[34] Stevens, Stanley Smith, Volkmann, John, and Newman, Edwin Broomell. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

[35] Tang, Yansong, Ni, Zanlin, Zhou, Jiahuan, Zhang, Danyang, Lu, Jiwen, Wu, Ying, and Zhou, Jie. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9839–9848, 2020.

[36] Traunmüller, Hartmut and Eriksson, Anders. The frequency range of the voice fundamental in the speech of male and female adults. *Unpublished manuscript*, 11, 1995.

[37] Valstar, Michel, Gratch, Jonathan, Schuller, Björn, Ringeval, Fabien, Lalanne, Denis, Torres Torres, Mercedes, Scherer, Stefan, Stratou, Giota, Cowie, Roddy, and Pantic, Maja. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10, 2016.

[38] Williamson, James R, Godoy, Elizabeth, Cha, Miriam, Schwarzentruber, Adrianne, Khorrami, Pooya, Gwon, Youngjune, Kung, Hsiang-Tsung, Dagli, Charlie, and Quatieri, Thomas F. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18, 2016.

[39] Yan, Sijie, Xiong, Yuanjun, and Lin, Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[40] Zhao, Yan, Liang, Zhenlin, Du, Jing, Zhang, Li, Liu, Chengyu, and Zhao, Li. Multi-head attention-based long short-term memory for depression detection from speech. *Frontiers in Neurorobotics*, page 111, 2021.