Universiteit
Leiden

# Master Computer Science

## DPD (DePression Detection) Net: A Deep Neural Network for Multimodal Depression Detection

| | |
|---|---|
| Name: | Manlu He |
| Student ID: | s2407590 |
| Date: | 05/02/2024 |
| Specialisation: | Artificial Intelligence |
| 1st supervisor: | Dr. Erwin M. Bakker |
| 2nd supervisor: | Prof. Dr. Michael S. Lew |

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

# Abstract

Depression is one of the most prevalent mental conditions which could impair people's productivity and lead to severe consequences. The diagnosis of this disease is complex as it often relies on a physician's subjective interview-based screening. The aim of our work is to propose deep learning models for automatic depression detection by using different data modalities, which could assist the diagnosis of depression. Current works on automatic depression detection mostly are trained and tested on a single dataset, which might lack robustness, flexibility and scalability. To alleviate this problem, we design a novel Graph Neural Network-enhanced Transformer model named DePressionDetect Net (DPD Net) that leverages textual, audio and visual features and can work under two different application settings: the clinical setting and the social media setting. We also propose a model named DePressionDetect-with-EEG Net (DPD-E Net) to incorporate Electroencephalography (EEG) signals and speech data for depression detection. Experiments across four benchmark datasets show that DPD Net and DPD-E Net can outperform the state-of-the-art models on three datasets (i.e., E-DAIC dataset, Twitter depression dataset and MODMA dataset), and achieve competitive performance on the fourth one (i.e., D-vlog dataset).

# Contents

# Chapter 1

# Introduction

Depression (also known as major depressive disorder) is a prevalent mental disorder with serious impact on people's personal life and the society. According to World Health Organization (WHO), there are approximately 280 million people in the world that suffer from depression and it is ranked as the fourth leading cause of death among people aged 15-29[1]. Researchers predict it to be the second leading cause of burden of disease by 2030 [1]. The economic impact of depression was estimated at €92 billion annually in the European Economic Area (EEA) [2]. Additionally, the COVID-19 pandemic caused a 27·6% increase in cases of major depressive disorders globally [3].

The most common approach of depression screening is based on physician-administered interview using questionnaires such as the Physical Health Questionnaire Depression Scale (PHQ) [4]. This type of screening highly relies on a physicians' subjective interpretation [5]. Also, some patients are reluctant to share their honest thoughts and talk about their symptoms during the screening interviews as they are ashamed of the stigma attached to depression.

Researchers have shown increasing interest in machine learning-based automatic depression detection using behavioural cues such as facial activity, gesturing, head movements and speech as studies have shown that they are strongly correlated with depression [4]. By integrating different cues, information from single modalities can complement each other, and the fused information has the potential of revealing underlining depression-related patterns. Automatic depression detection remains a difficult task for researchers given the following challenges. Firstly, public available multi-modal data of depressed individuals in clinical settings is limited. Secondly, a thorough depression interview session is usually long, so extracting useful context information from long sequences of audiovisual data for depression detection can be challenging [6]. Furthermore, existing studies solve the problem of multi-modal depression detection in one single setting, e.g., the clinical setting or the social media setting, and mostly are validated on a single dataset, which might lack

[1] World Health Organization. https://www.who.int/en/news-room/fact-sheets/detail/depression

robustness, flexibility and scalability.

Based on the work of Joshi et al. [7] named COGMEN, where the Transformer architecture and Graph Neural Networks (GNNs) are leveraged for modeling global dependencies and local dependencies respectively, we introduce a GNN-enhanced Transformer model named DePressionDetect Net (DPD Net) with novel convolution modules for automatic depression detection, which can be applied to the audio, speech and textual modality in different settings and work across different datasets while addressing the challenges mentioned above. Also, we design an extended version of the DPD Net called DePressionDetect with EEG Net (DPD-E Net) to incorporate Electroencephalography (EEG) data for depression detection. In this study, we investigate the problem of automatic depression detection in a multi-modal setting. Given audio, visual cues, textual cues and EEG signals, our proposed model should be able to predict the severity of depression in terms of Physical Health Questionnaire Depression Scale (PHQ) scores or predict if the subject is experiencing depression. Our main contributions are:

1. A novel GNN-enhanced Transformer model named DePressionDetect Net (DPD Net) is proposed based on the work of [7], which is originally adopted for muti-modal emotion recognition, we upgrade the model so it can be better adapted to muti-modal depression detection using audio, visual and textual cues. To be more specific, we change the structure of the previously proposed model and use it as a multi-modal encoder module, and design a novel unimodal encoder module which consists of conv-based sub-modules for encoding each single modality along with a detection module.

2. An ensemble model named DePressionDetect with EEG Net (DPD-E Net) is proposed for depression detection in EEG signal and speech data, which takes both the spatial and temporal information of the EEG signal into consideration.

3. We perform feature engineering and evaluate our proposed models on four depression datasets, which are under two different application settings: the clinical setting and the social media setting. Comparisons to other baseline methods show that DPD Net and DPD-E Net can outperform the state-of-the-art models on three datasets (i.e., E-DAIC dataset, Twitter depression dataset and MODMA dataset), and achieve competitive results on the rest one (i.e., D-vlog dataset).

4. Extensive experiments together with ablation studies are conducted to investigate the impact of the representation methods, different modules of our proposed model and the effect of the different modalities.

The remainder of this paper is organized as follows. Chapter 2 summarizes the related works. Chapter 3 introduces fundamentals used in this work. Chapter 4 describes the details of the baseline methods. Chapter 5 presents the dataset, the procedure of data preprocessing and feature extraction. Chapter 6 describes the details of our proposed models, DPD Net and DPD-E Net. Experiments and the results are shown in Chapter 7. Chapter 8 concludes the work of the thesis.

# Chapter 2

# Related Work

This chapter gives an overview of the works on automatic depression detection which are relevant for our study.

**Depression Detection from Speech** In current literature, many studies in the domain of automatic depression detection from speech utilize Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) as the backbone of their proposed models. Adrián et al. [8] design an ensemble model which consists of three stacked CNN base learners to balance bias and variance to perform depression detection from speech on the Distress Analysis Interview Corpus-Wizard of Oz database (DAIC-WOZ) [9]. This database is considered as one of the benchmark databases in the domain of multi-modal depression detection under the clinical setting, and its newly extended version called E-DAIC dataset [4] that contains more subjects is used in our study. Zhao et al. [10] introduce a bidirectional long short term memory (BiLSTM) model with a hierarchical attention mechanism to differentiate depressed individuals from healthy controls using various frame-level audio features. The application of Transformer-based model is presented in the work of [11]. Here a model termed transformer-CNN-CNN (TCC) which integrates a parallel-CNN module to capture local information with a 4-layer transformer to capture long term sequential knowledge is proposed, and achieves state-of-the-art performance on the DAIC-WOZ dataset.

**Depression Detection from Visual Cues** Even though head movements and gesturing are also considered as discriminative cues for depression detection [4], most of the researches focus on the utilization of facial activities. Zhou et al. [12] propose a model named Multi-Region DepressNet to predict depression levels using sequential images of facial regions cropped from interview videos. Each single DepressNet module is based on a pre-trained ResNet50 [13] network with modifications on the prediction block. By stacking four DepressNet modules aimed for different facial regions, features learnt from each region are combined to make the final prediction. Wang et al. [14] proposed a BiLSTM-based framework for classifying depressed and non-depressed individuals using facial landmarks. They model this task as a multiple instance learning problem by considering each complete video file of a patient during screening as a bag and each sliced video segment as an instance [14]. Given that the dynamics of facial expression could play an essential role in depression detection, Melo et al. [15] design a novel encoding approach that performs temporal pooling on sequential video images. The raw images along with the encoded images are input to two separate CNN-based regressors for prediction and scores are fused by taking the average [15]. This model is able to produce state-of-the-art results on two video-based depression datasets.

The use of visual related bio-signals is presented in the work of Casado et al. [16], where they investigate the effectiveness of remote photoplethysmography signals directly extracted from facial videos in depression detection. Using a random forest model, the proposed method is evaluated on two public available datasets, and achieves similar results compared to some deep learning approaches, such as the two-stream CNN regressor [15] mentioned above.

**Depression Detection from Text** As textual data is convenient to be collected and normally could be accessible after proper de-identification, it is widely used in the domain of automatic depression detection. Ansari et al. [17] develop a hybrid model to recognize depressed users on social media platforms using the user-generated textual content. The model consists of a logistic regression classifier that takes four types of lexicon as input, and an attention-enhanced LSTM that takes word embeddings as input. Zhang et al. [18] design a model that integrates a modified Robustly optimized BERT approach (RoBERTa) model [19] with a BiLSTM model. Before feeding text data into the model, a random oversampling technique is adopted to overcome the data imbalance issue. Evaluations on two clinical datasets of different languages indicate the model's potential. Based on a finding that the Bidirectional Encoder Representations from Transformers (BERT) [20] component brings most of the performance gain of a depression detection model called Audio-Assisted BERT (Audibert) [21], Saskia et al. [22] investigate the effectiveness of a BERT ensemble model comprised of a BERT base model along with its two varaints named RoBERTa [19] and DistilBERT [23], using clinical textual contents. Experiments show that the ensemble can improve depression detection performance compared to individual models and introduce stability.

**Multi-modal Depression Detection** Since depression is a mental condition of complex patterns, most of the studies utilize multi-modal information to fully exploit each modality's predictive power. To deal with the long-interview problem we mentioned in Chapter 1, Gong et al. [24] propose a novel approach for multi-modal depression detection based on the topics of the clinical interview recording. They extract the topic of each answer of the interview questions and obtain the corresponding textual, visual and acoustic features of the answer. Different features of the same topic are aggregated to construct new feature vectors so the sequence length of each sample is decreased to the total number of topics [24]. Makiuchi et al. [25] introduce three different CNN-LSTM unimodal models for encoding visual, textual and audio cues. Then, the learnt features from each modality are concatenated for a linear regressor that predicts the depression score.

Transformer-based models gained their popularity in multi-modal depression detection given their high performance in the field of Natural Language Processing (NLP) and Computer Vision (CV). Audibert [21] designed by Ermal et al. is a depression detection model that takes two streams of input. A textual stream is sent to pre-trained BERT while an audio stream is feed into pre-trained audio networks such as Wav2vec [26] and SincNet [27], both of which are followed by a BiLSTM and their output are aggregated for depression prediction. In [6], authors extract Mel-Frequency Cepstrum Coeffiecient (MFCCs) and Facial Action Units (FAUs) from depression screening videos, and design a Transformer-based model which leverages these two modality and performs multi-task learning by treating classification as an auxiliary task of the main regression task.

Even though text, visual cues and audio cues are the most commonly used modalities for depression detection, estimating depression from Electroencephalography (EEG) data has become a promising research direction, as EEG signals of high temporal resolution can well capture complex brain activities and turn out to be suitable for depression-related research [28]. The Multi-modal Open Dataset for Mental-disorder Analysis (MODMA) is introduced by [29] to boost research in depression detection using physiological data. The dataset comprises both speech data and EEG data collected under clinical setting, but most current studies only use one modality.

As the depression data under clinical settings is limited, depression datasets curated by collecting social media contents are investigated by researchers to discover potential depression-related patterns in daily life. The D-vlog Dataset [30] is an audio-visual depression dataset collected from daily video blogs, and the Twitter Depression Dataset [31] is build by extracting user posts which contain tweets accompanied with images.

Our work differs from all the previous works in terms of the method and the scope of our research. By introducing a novel unimodal encoder module, a detection module, and modifying the architecture of the network [7] as a multi-modal encoder module, our proposed model is the first work to our knowledge that uses a GNNs-enhanced transformer in multi-modal depression detection. Also, we use four aforementioned benchmark datasets including the E-DAIC dataset [4], the D-vlog Dataset [30], the Twitter depression dataset [31] and the MODMA dataset [29], which cover different settings for a comprehensive study in this domain, where the proposed models have the potential to be applied to real scenes. All the datasets used in this work are described in detail in Chapter 5.

# Chapter 3

# Fundamentals

This chapter gives descriptions for all the fundamentals that are related to our project.

**Evaluation Metrics** In general, our proposed models tackle two types of problems: the classification problem and the regression problem. We leverage the following metrics to evaluate the performance of the models.

For the E-DAIC dataset, we use the Concordance Correlation Coefficient (CCC) score to evaluate the model's regression performance. This metric is considered as the only criteria for ranking participants' methods of the depression detection challenge proposed by the original E-DAIC dataset paper, and is the most widely used metric for depression detection research on the E-DAIC dataset, as it is a metric which perfectly takes both precision and accuracy into consideration in one formula and is robust to scale variance [4]. The CCC score measures the correlation between the prediction and the ground truth using the following equation:

$$CCC_{score} = \frac{2C_{py}}{C_p^2 + C_y^2 + (\bar{p} - \bar{y})^2} \tag{1}$$

Note that a depression score ranges from 0 to 24. Here $\bar{p}$ is the mean value of the predicted depression scores and $\bar{y}$ is the mean value of the true depression scores. $C_p$ and $C_y$ are the variance of the predicted and the real depression scores, respectively. $C_{py}$ is the covariance of the predicted and the real depression scores. The $CCC_{score}$ ranges from -1 to 1, where 1 indicates the highest correlation between the predictions and the ground truths.

Also, we report Root Mean Square Error (RMSE) as an auxiliary metric, which is calculated using Equation 2. Here $N$ is the number of samples, $p_i$ and $y_i$ denote the predicted depression score and the real depression score of sample $i$, respectively.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(p_i - y_i)^2}{N}}, \tag{2}$$

For the classification tasks on the Twitter Depression dataset, D-vlog dataset and MODMA dataset, we use precision, recall and F1-score as the metrics. Precision represents the percentage of all the positive labels that are correctly assigned as positive by the model. Let $TP$ denote the number of true positives, and $FP$ denote the number of false positives, then precision is computed as follows:

$$precision = \frac{TP}{TP + FP} \tag{3}$$

Recall is a metric to measure, for all the positive samples, what percentage of them are actually found by the model, and is formulated as:

$$recall = \frac{TP}{TP + FN} \tag{4}$$

The F1-score takes both the precision and recall into consideration using the following equation:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

# Chapter 4

# Baselines

This chapter introduces the baseline methods that our proposed models are compared to. Different baseline methods are selected for comparison on four aforementioned benchmark datasets including the E-DAIC dataset [4], the Twitter depression dataset [31], the D-vlog Dataset [30] and the MODMA dataset [29]. All of the baselines are transformer-based models and produce the state-of-the-art results on each of the datasets, respectively.

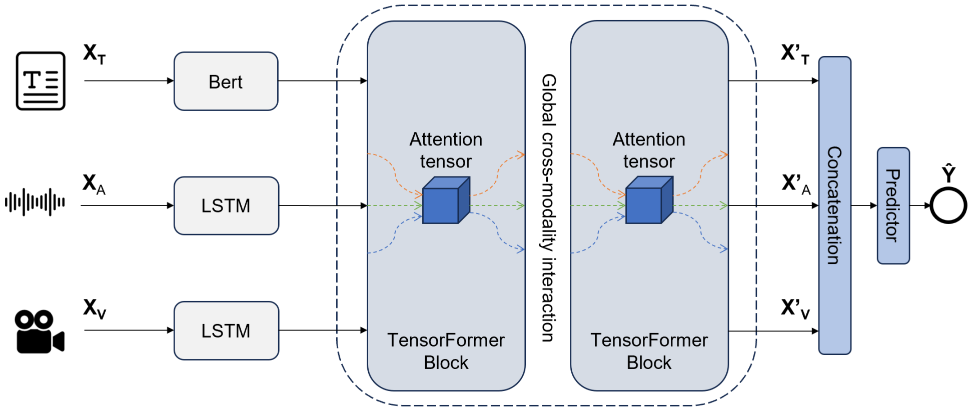


**Figure 4.1:** The overall architecture of TensorFormer.

**TensorFormer** In the work of [32], the authors propose a tensor-based multimodal Transformer called TensorFormer for depression detection using text, audio and videos, which is the state-of-the-art method on the E-DAIC dataset. The general overview of the model is shown in Figure 4.1. Here $X_T$, $X_A$, and $X_V$ denote the input text, audio features, and visual features, respectively. TensorFormer starts with a pre-trained BERT model that encodes texts, and two bidirectional long short term memory (BiLSTM) models to encode the input audio and visual features. The obtained features from these three modalities are then processed by the TensorFormer block, which performs global cross-modality interaction through the attention tensor. The attention tensor is calculated by the Cartesian product of the features from three modalities, and is used for computing the weighted summary of each modality which are able to retain complementary information from the involved modalities [32]. Subsequently, new representations of each of the modalities, denoted by $X'_T$, $X'_A$, and $X'_V$, are forwarded to their corresponding feed forward modules with residual connections, and the further encoded embeddings from each modalities are finally concatenated to make the prediction denoted by $\hat{Y}$. By stacking 2 layers of the TensorFormer blocks, their experiments have shown that TensorFormer not only achieves state-of-the-art performance on multi-modal depression detection but also produces promising results on multi-modal sentiment analysis.

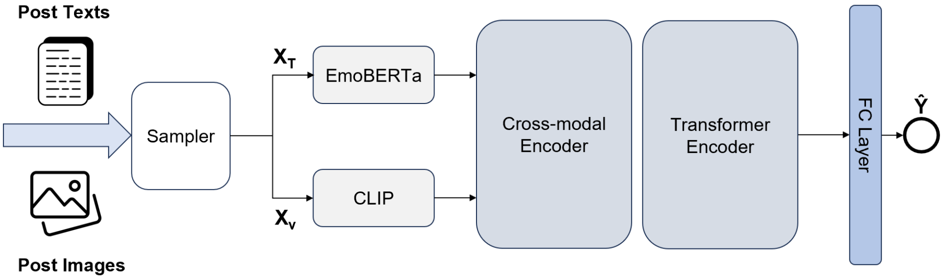


**Figure 4.2:** The overall architecture of TM Transformer.

**TM Transformer** For the Twitter depression dataset, the Time-enriched Multimodal Transformer proposed by [33] achieves state-of-the-art performance in detecting depression from twitter users based on textual contents with images that are posted together with the text. Figure 4.2 depicts the architecture. Firstly, pre-trained models including CLIP [34] and EmoBERTa [35] are leveraged for encoding images and texts, denoted by $X_T$ and $X_V$, respectively. Then, the features learnt from each modality along with two types of positional embeddings are send to the cross-modality attention module that can capture informative patterns through modalities' interaction and the further encoded embeddings are passed to a classic transformer encoder followed by a fully connected layer to make the final prediction [33]. They propose three models based on the used positional embeddings and the post sampling strategies: VanillaTransformer uses the classic learnt positional encodings with sub-sequence sampling; SetTransformer leverages zero positional encoding with random sampling of user posts; Time2VecTransformer uses time-enriched positional embeddings with sub-sequence sampling. Their experiments show that Time2Vec Transformer is able to outperform the other two models, which illustrates the effectiveness of time2vec positional embeddings. In our work, we compare our DPD Net with all the three models.

**STST** For the D-vlog dataset, we use the spatio-temporal squeezed transformer (STST) shown in Figure 4.3 as the baseline, which adopts cross-attention mechanism to extract spatio-temporal features for depression detection [36]. After normalizing and transforming the audio and video data into the same shape, two layers of STST blocks are applied to the pre-processed features to learn spatio-temporal patterns from each modality using a novel cross-attention method. The STST block works by passing the obtained values (V) to an additional convolutional-based block, comprising three 1D convolutional layers and a ReLU activation, to model spatial information, and the adjusted values (V) are used along with the obtained query (Q) and keys (K) to perform the cross-attention that models interactions among different modalities [36]. Finally, the updated audio features and visual features are concatenated and forwarded to a multi-layer perceptron (MLP) to make the prediction.

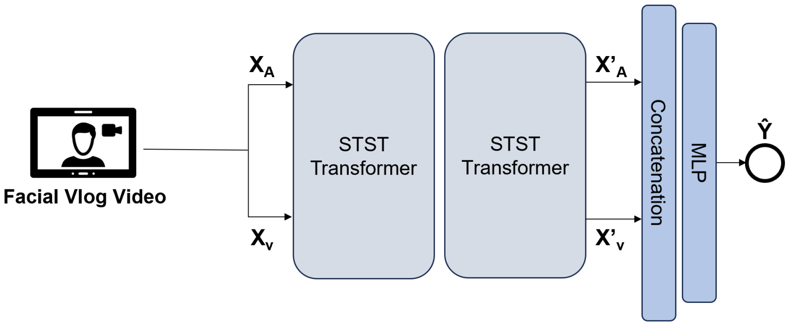


**Figure 4.3:** The overall architecture of spatio-temporal squeezed transformer (STST).

**ES Vision Transformer** We use the state-of-the-art vision transformer-based model proposed by [37] as the baseline method for the Multi-modal Open Dataset for Mental-disorder Analysis (MODMA) dataset. As can be seen from Figure 4.4, this model is comprised of three main modules including a CNN-LSTM module for encoding the raw EEG signal denoted by $X_{E1}$, a vision transformer module to learn features from EEG spectrogram denoted by $X_{E2}$, and another vision transformer module for extracting features from speech spectrogram denoted by $X_A$. The learnt features from each module are fused to predict the depression labels through a classifying module consisting of two fully connected layers and a ReLU activation.

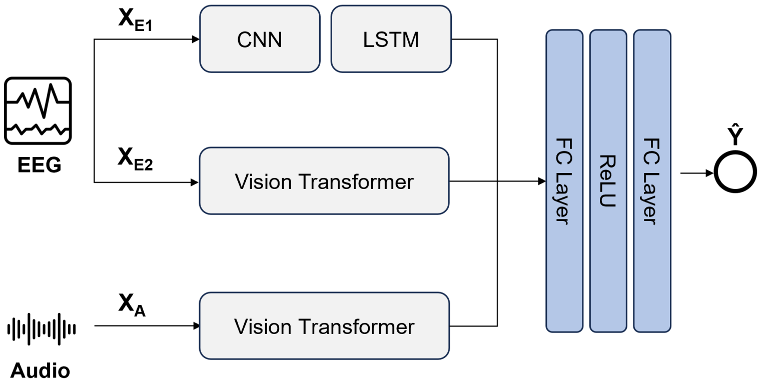


**Figure 4.4:** The overall architecture of ES Vision Transformer.

# Chapter 5

# Datasets

For this study, we use four different datasets in the multi-modal depression detection domain to validate our proposed methods. The first dataset we use is the **E-DAIC Database** [4], which is a depression dataset of multi-modal data collected from clinical interviews. As multi-modal depression-related data in the clinical setting is rarely publicly available, we use the **Twitter Depression Dataset** [31] and the **D-vlog Dataset** [30] which are collected from social media to further evaluate our method and explore the domain of multi-modal depression detection. Speech, visual, and textual modalities are involved in these three datasets. The **MODMA Dataset** [29], a Chinese multi-modal depression dataset which contains audio data and EEG signals is used to investigate DPD-E Net's effectiveness. Detailed descriptions of each dataset, the procedure of data pre-processing and feature extraction are presented in this chapter.

## 5.1 E-DAIC Database

### Dataset Description

Extended DAIC Database (E-DAIC) [4] is a multi-modal depression dataset of clinical interview recordings in which all interviews are conducted by a virtual avatar who asks general questions such as personal information, or questions related to the Physical Health Questionnaire Depression Scale (PHQ) [5]. Audio, video and the transcript of each participant's interview are recorded in the dataset along with their final PHQ scores. The range of the PHQ score is 0 to 24 and higher scores indicate more severe depression [4]. Table 5.1 gives an example of an interview transcript from a participant with the highest PHQ score.

The dataset consists of 275 samples of interview sessions with 40.5 hours recordings in total, and is split into a training set, validation set and test set. Some statistics are presented in Table 5.2. The average duration of an interview session is about 16 minutes on average, which poses challenge for our prediction task as the estimation needs to be made from the

|  | Sample Questions | Sample Answers |
|---|---|---|
| 0 | Where are you from originally? | Los Angeles California. |
| 1 | Who is someone that has been a positive influence in your life? | My teachers. |
| 2 | Tell me more about that. | Spiritual teachers that I find a lot of guidance from. |
| 3 | What are some things that make you really mad? | The situation with my life right now. |
| 4 | Can you tell me more about that? | I can't find a job. I applied from anywhere and everywhere, from entry-level to management, anywhere in between. |
| 5 | How easy is it for you to get a good night sleep? | It isn't easy. |
| 6 | Does it happen quite often? | Yes. It happens very often. |
| 7 | When was the last time that happened? | Last night I couldn't sleep. |
| 8 | Why? | Just thinking about my situation. Car payment was due yesterday. I just don't know if I have what it takes to continue to do. |
| ... | ... | ... |

**Table 5.1:** An interview transcript example of a depressed participant.

whole recording of the interview while the duration of a prediction sample (i.e., an audio or video segment used for prediction) is so much shorter (seconds-long) for other multi-modal tasks such as emotion recognition [24]. Also, the number of samples is limited, but E-DAIC is the only public available depression dataset with data of three modality (audio, video, text) and is considered as the benchmark dataset in this domain [4].

|  | Samples | Duration (hrs) | Avg Duration (mins) |
|---|---|---|---|
| Training | 163 | 40.5 | 16.0 |
| Development | 56 | 14.8 | 15.8 |
| Test | 56 | 14.9 | 15.9 |
| All | 275 | 70.2 | 15.9 |

**Table 5.2:** Statistics of the training, development and test set in E-DAIC Dataset.

### Data Pre-processing

The original E-DAIC dataset contains raw audio data with a sampling rate of 16000 Hz and a resolution of 16 bits, the transcript of the audio is given in csv format along with the corresponding timeframe of each utterance. Raw video files are not released due to privacy concerns, while Facial Action Units (FAUs), gaze and pose positions of the participants extracted from interview videos are provided. We pre-process the dataset in the following steps.

1. Starting with the transcripts, we remove all the questions by pattern matching, as we find that all questions follow certain formats at the utterance beginning, such as

'can/have/do/did/are you', 'what are/is/do/would/got', or contain certain phrases such as 'last time', 'tell me'. By deleting the question part, we alleviate the long sequence problem without undermining the effective information that should be retained in the data, which is illustrated by the experimental results in Section 7.2.3.

2. There are errors of the given timeframes of the transcripts, such as the starting time of certain utterance being earlier than the end time of its previous utterance. We track these errors and fix them by manually checking the raw audio files.

3. After obtaining the new transcripts, we slice the audio file into segments of utterances using the given timeframes and keep their corresponding visual descriptors. The provided visual descriptors are discussed in more detail in Section 5.5.

## 5.2 Twitter Depression Dataset

### Dataset Description

With the surge of social media, people often share their daily lives and opinions using platforms such as Twitter in a tweet accompanied with a related image or a video. This opens up a new direction for automatic depression detection, as a user's profile can be represented by their posted tweets and could possibly indicate their mental state. With additional visual cues from the posted images, richer information for depression estimation becomes available.

| | User | Text | Text + Images |
|---|---|---|---|
| Depressed | 1,402 | 232,895 | 22,195 |
| Non - Depressed | 1,402 | 879,025 | 64,359 |
| All | 2,804 | 1,111,920 | 86,554 |

**Table 5.3:** Statistics of the Twitter Depression Dataset.

The Twitter Depression Dataset [31] is build upon another dataset which only contains textual modality. Authors curate the new dataset by extracting extra tweets accompanied with images using user IDs obtained from the text-based dataset [31]. This dataset is the most widely used multi-modal depression detection dataset in the social media setting. In total, it includes 2804 user samples with approximately 110M tweets. Users are identified as depressed individuals or healthy controls, and the distribution of these two classes is balanced, which can be seen from Table 5.3. Both of the classes have hundreds of thousands of tweets but only around 7 to 10 percent of the tweets are followed by images. Also, the non-depressed class has on average more tweets collected per user.

We present an example of a depressed user in Figure 5.1. It is worth noting that in this case, the textual information is enhanced by the corresponding images. Especially, for the first tweet, the textual content *'What am I doing?'* is a plain self-question which seems to not indicate any depression tendency. But with its corresponding image, some hidden mental condition information might be revealed.

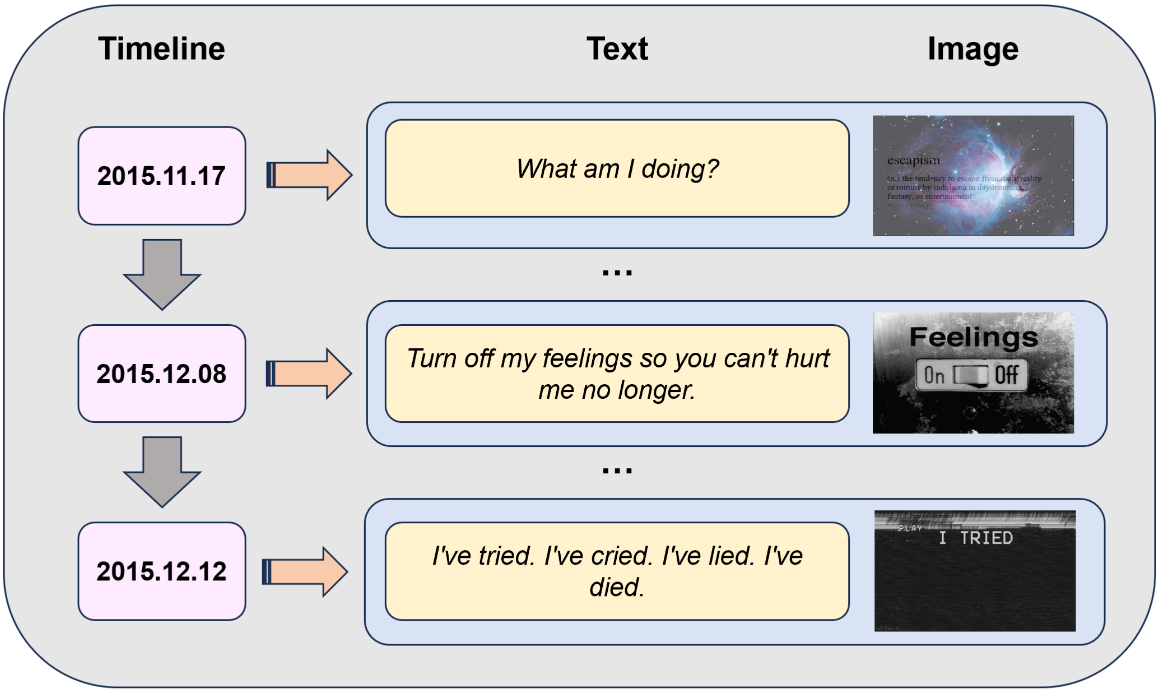

Timeline
Text
Image
2015.11.17
What am I doing?
escapism
(n.) the tendency to escape from daily reality or routine by indulging in daydreaming, fantasy, or entertainment
...
2015.12.08
Turn off my feelings so you can't hurt me no longer.
Feelings
On Off
...
2015.12.12
I've tried. I've cried. I've lied. I've died.
PLAY
I TRIED

**Figure 5.1:** An example of a depressed user in the Twitter Depression Dataset.

### Data Pre-processing

Each tweet sample in the Twitter Depression Dataset is presented as key-value pairs such as posted time, tweet ID, text and user interactions. Images are named using their corresponding tweet IDs. As the scope of this study is multi-modal depression detection and we would like to explore the performance of our model on the text and visual modalities, we only keep those tweets which were posted with images as our experiment data. We pre-process the data as follows:

1. Only relevant data fields are kept, including post time, tweet ID and text. There are errors of the image files as some of them are zero-byte. We remove these empty images and their tweets, and sort each user's tweets in time ascending order.

2. Tweet textual content normally contains various types of noise so we clean each tweet by removing links, emojis, user mentions and special characters. The hashtag symbols are deleted but the content followed by are kept, as it is quite common for social

media users to use hashtags to summarize their posts. Then, we convert all tweets to lowercase, and remove all the empty tweets generated after text cleaning.

3. For all the images, we check them if they are RGB files, and transform them to RGB if they have an additional alpha channel.

## 5.3 D-vlog Dataset

### Dataset Description

Video-sharing platforms such as Youtube offer opportunities for depression detection in daily lives as video blogs posted by users might imply their psychological state. The D-vlog Dataset [30] is an audio-visual depression dataset consisting of daily video blogs from Youtube. The selected videos have the subject speaking directly to the camera to make sure each subject's facial activities can be leveraged for depression estimation [30].

| | **Train** | **Validation** | **Test** |
|---|---|---|---|
| Depressed | 375 | 57 | 123 |
| Non - Depressed | 272 | 45 | 89 |
| All | 647 | 102 | 212 |

**Table 5.4:** Statistics of the D-vlog Dataset.

The dataset contains 961 samples of 816 subjects, with each sample having an average length of around 10 minutes. As can be seen from Table 5.4, the dataset is split into the training set, validation set and test set with a ratio of approximately 7:1:2, and the number of depressed individuals is slightly larger than that of the non-depressed.

### Data Pre-processing

Videos are segmented into 1-second fragments. Their corresponding 25 low-level acoustic descriptors and 68 facial landmarks are extracted and concatenated to form the audio and visual descriptors. For ethical concerns, the dataset only provides these descriptors as feature vectors instead of the raw videos and audios [30]. We check the length of the given audio and visual features to see if they are synchronized. For those of unmatched lengths, we simply pad the shorter features with zeroes.

## 5.4 MODMA Dataset

### Dataset Description

Recent studies often use EEG data for automatic depression detection in a unimodal way. To make use of information from additional modalities, the Multi-modal Open Dataset for Mental-disorder Analysis (MODMA) [29] is used in this project, which is the only available open dataset which comprises both speech data and EEG data. The speech data are recorded during the interviews in which each subject is asked 18 depression-related questions, and only the audio of the answering part are kept. All the interviews are conducted in Chinese. EEG signals are collected under resting-state for 5 minutes using a 128-channel HydroCel Geodesic Sensor Net with a sampling frequency of 250Hz [29].

| | **Speech** | **EEG** | **Gender (F/M)** |
|---|---|---|---|
| Depressed | 24 | 23 | 18/29 |
| Non - Depressed | 29 | 29 | 18/40 |
| All | 53 | 52 | 36/69 |

**Table 5.5:** Statistics of the MODMA Dataset.

In total, 53 subjects are included in the speech data and 52 participants' EEG signals are collected. According to Table 5.5, the dataset is quite sparse but balanced. Also, there are more male subjects for the healthy control group and the depression group.

### Data Pre-processing

After checking the unique subject IDs assigned to the participants, we find that not all participants participate in both the interview test and the EEG test. Most of the works in detection depression with EEG data use single EEG modality, which might not able to fully uncover the depression-related patterns behind. As the goal of this project is to estimate depression from multi-modal data, we only kept those participants who are involved in both tests, which leads to a total of 33 subjects. Then, we perform data pre-processing through the following steps:

1. For the EEG data, we use finite impulse response (FIR) filter of 0.5–50 Hz as this is the frequency range where most of the depression-related signals are located [38]. Then, the processed EEG signals are re-referenced to the average electrode, which is a common practice in EEG signal processing to remove background noise [39]. Lastly, the processed EEG signals are segmented to 8-second epochs.

2. For the speech data, the original audio data is provided after being segmented into audio clips, where each audio clip represents a corresponding answer for the interview question. Unlike the E-DAIC dataset which is also collected under clinical interviews, transcripts of the interview are not provided in the MODMA dataset. We manually transcribe the audio recordings into Chinese text to include the textual modality.

## 5.5 Feature Extraction

### Audio Features

Low-level audio features including Mel-Frequency Cepstrum Coeffiecient (MFCCs) and Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) are extracted for the E-DAIC and the MODMA dataset. MFCCs are popular features in speech-related applications as it is using filter bank inspired by the human's perception of the speech signal [40]. eGeMAPS is a feature set consisting of frequency/energy related parameters, spectral parameters and their functionals [41]. The reason for using these features is that, the original E-DAIC dataset also provides other baseline deep-learnt features. However, after some preliminary experiments with both the low-level features and deep-learnt features, we find that using low-level features can lead to better performance with DPD Net. We extract 40 MFCCs using Librosa[1] and 88 eGeMAPS features using Opensmile[2] for each utterance (answer), which are then concatenated into a 128-dimensional audio feature. All the 88 parameters are given and explained in the original eGeMAPS paper [41]. For the D-vlog dataset, the provided low-level acoustic descriptors including loudness, MFCCs, spectral flux, etc., which constitute 25-dimensional vectors, are used as audio features.

### Visual Features

For the E-DAIC dataset, we use 18 provided Facial Action Units (FAUs) which are quantified by intensity as visual features. They describe the activities of certain facial muscles such as the raise of upper lid (AU5) and wrinkle nose (AU9), which can be seen from Figure 5.2. We also use the Gaze position and pose position as visual features, which present the direction of gaze and the position and orientation of the head respectively [4]. The combination of these three types of visual descriptors leads to a 49-dimensional visual feature. For the D-vlog dataset, 68 facial landmarks which presents the locations of face features in the form of x and y coordinates, are used as visual features.

For the Twitter depression dataset, instead of learning image representations from scratch, we leverage the pre-trained CLIP [34] model as our visual feature extractor. CLIP proposed

[1] https://librosa.org

[2] https://www.audeering.com/research/opensmile

[3] https://github.com/TadasBaltrusaitis/OpenFace/wiki/Action-Units

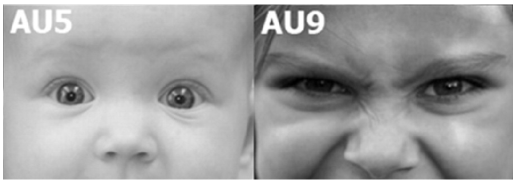


**Figure 5.2:** Example of FAUs that encode facial activities.[3]

by OpenAI is a multi-modal model which is trained on 400M image-text pairs crawled from internet by predicting if the textual content is aligned with an image [34]. Hugging Face toolkit [42] is used for extracting features from raw tweet images. Firstly, we use 'AutoProcessor' to convert images to the correct input format, then 768-dimensional visual features are obtained from Hugging Face pre-trained models 'clip-vit-base-patch16'.

### Textual Features

We adopt a pre-trained language models MpNET [43] for encoding the texts. By experimenting with pre-training settings such as batch size, masking method and the choice of input, MpNET is proposed as an upgraded version of BERT [20] using the best settings found in their experiments [43]. To encode the English text from E-DAIC dataset and Twitter depression dataset, we use the Hugging Face pre-trained model 'all-mpnet-base-v2'. As for the Chinese text in MODMA dataset, the multilingual-model 'paraphrase-multilingual-mpnet-base-v2' is utilized. These two models are selected as text encoders based on our preliminary experiments.

### EEG Features

In this work, 3-stream EEG data are used for estimating depression, which consist of two types of EEG features: temporal features presented by a combination of different linear and non-linear features, and spatial features presented by Brain Functional Networks (BFNs).
From the 128 available electrodes, we select 16 main electrodes (channels) in Figure 5.3, which are highlighted in blue. Because they are the most representative electrodes of EEG signal and have been proven to be effective in depression detection [44]. For each time step (i.e., 8-second epoch), we extract features such as activity, mobility, complexity, permutation entropy, spectral entropy, etc., which serve as input for the temporal EEG stream. In total, 29 linear and non-linear features listed in Table 5.6 are extracted for each channel and are finally flattened to 464-dimensional features (i.e., 29 * 16) as the input of temporal stream classifier.
Brain Functional Networks (BFNs) is a type of network that are constructed based on the the correlations among EEG channels. They describe the brain functional connectivity and have been widely used for brain disorder diagnosis [28]. We construct BFNs by considering

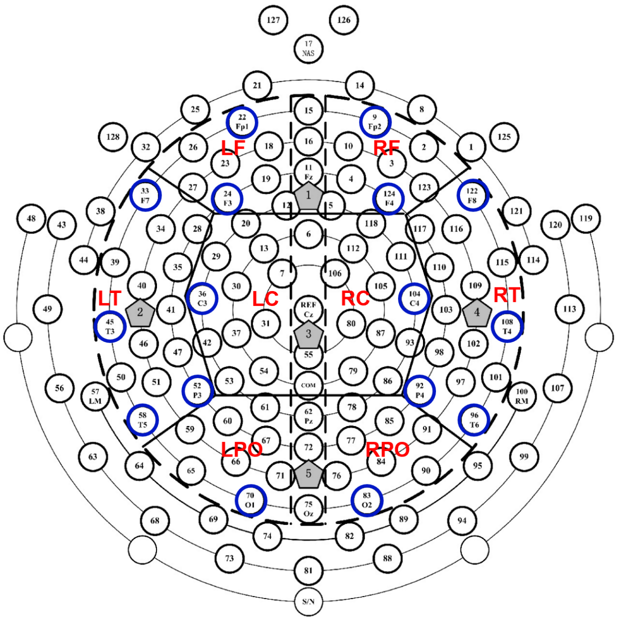


**Figure 5.3:** 128-channel EEG electrode positions and segmentation of brain regions: left frontal (LF), right frontal(RF), left temporal (LT), right temporal (RT), left central (LC), right central (RC), left parietal occipital (LPO) and right parietal occipital (RPO).

channels as nodes and using their Pearson Correlation Coefficient (PCC) to define the edges of the networks. For channels $x$ and $y$, the PCC is calculated as follows:

$$r = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{\sum (x - \overline{x})^2 \sum (y - \overline{y})^2}} \tag{6}$$

where $\overline{x}$ is the mean of the $x$ signal and $\overline{y}$ is the mean of the $y$ signal. The PCC value ranges between -1 and 1, with 0 meaning no correlation between these two signals. We take the absolute value, and set the threshold as 0.6 to define an edge. For the global spatial stream, we use the before-mentioned 16 main electrodes.

Brain regions can be divided into 8 sub-regions shown in Figure 5.3. We choose right temporal (RT), left central (LC) as the main sub-regions for exploring the local spatial information as it is found that channels belonging to these two sub-regions exhibit the most discriminative features between depressed and non-depressed subjects [45]. Then, all electrodes within each of the two sub-regions are leveraged to construct two separate BFNs as input to the local spatial streams.

| *Feature Name* | *Feature Name* |
|---|---|
| Kurtosis | Activity |
| Maximum of the Second Order Difference | Mobility |
| Mean of the Second Order Difference | Complexity |
| Maximum of the First Order Difference | Mean FFT Amplitude |
| Mean of the First Order Difference | Median FFT Amplitude |
| Coeffiecient of Variation | Min FFT Amplitude |
| Skewness | Max FFT Amplitude |
| Wavelet Approximate Mean | Mean FFT Amplitude for Alpha Band |
| Wavelet Approximate Standard Deviation | Mean FFT Amplitude for Beta Band |
| Wavelet Detailed Mean | Mean FFT Amplitude for Delta Band |
| Wavelet Detailed Standard Deviation | Mean FFT Amplitude for Theta Band |
| Wavelet Approximate Energy | Singular Value Decomposition Entropy |
| Wavelet Detailed Energy | Spectral Entropy |
| Wavelet Approximate Entropy | Permutation Entropy |
| Wavelet Detailed Entropy | |

**Table 5.6:** 29 linear and non-linear EEG features.

# Chapter 6

# Methodology

In this chapter, we present our two proposed models **DPD Net** and **DPD-E Net** together with their implementation details.

## 6.1 DPD Net

In this section, we introduce our novel model DePressionDetect Net (DPD Net) in detail. Figure 6.1 presents the overall structure of the DPD Net, which consists of three main components: a unimodal encoder module, a multi-modal encoder module and a detection module.

### 6.1.1 Model Overview

Given the text input denoted as $X_T$, audio descriptors as $X_A$ and visual descriptors as $X_V$, DPD Net starts with encoding each single modality. The proposed unimodal encoder module consists of a pre-trained MpNET model [43] for encoding the text into a 768-dimensional feature vector, and two novel convolution-based submodules that encode the audio and visual descriptors into 256-dimensional feature vectors, respectively.

After processing the input of each modality by their corresponding encoders, the features obtained from each modality are concatenated as a fused feature and fed into the multi-modal encoder module, which is based on the COGMEN model [7] originally proposed for multi-modal emotion recognition. To the best of our knowledge, there is no work in the domain of multi-modal depression detection that uses GNNs along with transformers, so our proposed model can be considered as a new attempt in this domain. The architecture of [7] is presented in Appendix A.

The Multi-modal encoder module begins with a transformer submodule including a transformer encoder and a linear layer, to produce a 100-dimensional multi-modal embedding. Based on this multi-modal representation, two types of homogeneous graphs are formed using past and future information, and are sent into a GNN submodule for further encoding. After this stage, another 100-dimensional representation is obtained and further concatenated with the embedding from the transformer submodule, and is sent to the detection module that we designed for making the final prediction. This is the general framework of the DPD Net, each of the submodules is described in the following sections.

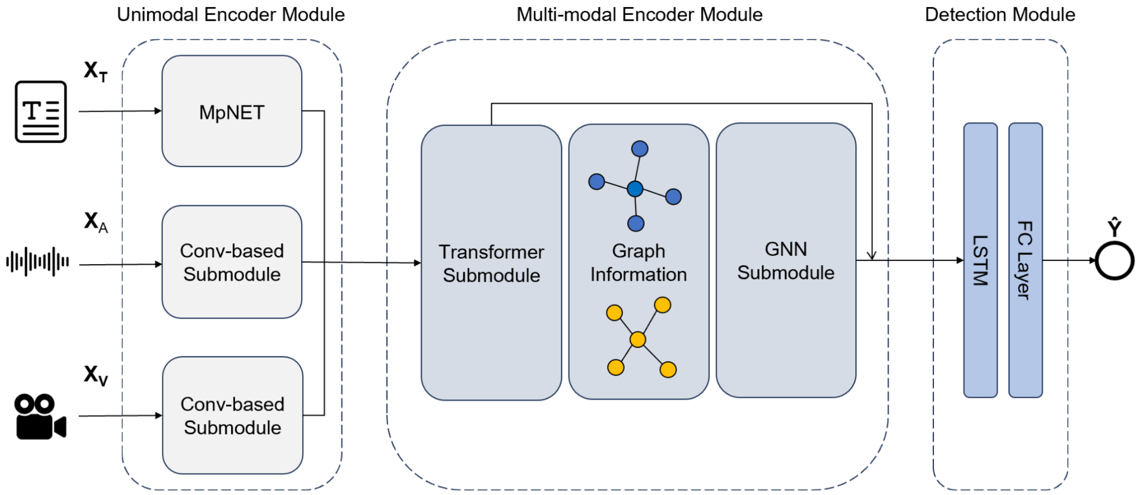


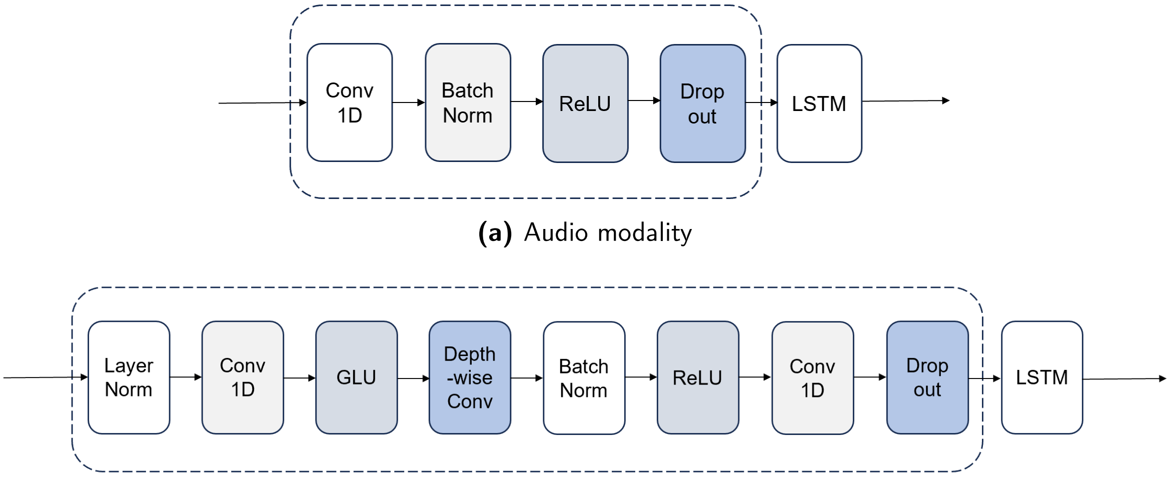
**Figure 6.1:** The overall architecture of the proposed DePressionDetect Net (DPD Net).

### 6.1.2 Unimodal Encoder Module

The unimodal encoder module is designed for learning representations for each single modality before the fusion of the modalities. MpNET [43] is adopted for encoding the textual content, as discussed in Section 5.5, and two conv-based submodules are proposed for encoding the audio and visual modality.

#### Audio Conv-based Submodule

As mentioned in the feature extraction Section 5.5, each audio segment is represented as a 128-dimensional feature vector containing MFCCs and eGeMAPs. For each interview session of a patient, we have $N$ answers, which are audio segments for each interview questions. For a vlog of a subject, we have sequential audio segments of 1-second with $N$ time steps. So the input to this audio encoding submodule is a vector with shape $(N, 128)$ for the E-DAIC dataset and the MODMA dataset. Since the provided audio descriptors of the D-vlog dataset are 25-dimensional vectors, the input shape in this case is $(N, 25)$.
To encode this audio information, we design a conv-based submodule for audio modality shown in Figure 6.2a. It starts with a 1D convlutional layer with 256 filters with kernel size 3 and stride 1, followed by a batch normalization, a rectified linear unit (ReLU) activation



**(a)** Audio modality

**(b)** Visual modality

**Figure 6.2:** The conv-based submodule for audio and visual modality.

and a dropout layer. The convolutional layer is designed for capturing local patterns of the input data by leveraging convlutional filters. In case of a 1D convlutional layer, each filter of certain kernel size slides over the input data in one axis with certain stride, and performs linear operations to produce feature maps in which essential patterns of the input data are encoded. In our preliminary experiments, we try to stack these layers (layers within the dash line cell in Figure 6.2a) with different kernel sizes and different number of filters, or using dilatation, but this simple structure gave the best performance. Then, the output of the dropout layer is fed into a 2-layer LSTM to capture longer dependencies of the audio modality.

### Visual Conv-based Submodule

For the visual encoder, we adopt another conv-based submodule inspired by the work of [46]. Note that the video features from the E-DAIC dataset are 49-dimensional, the posted images of the Twitter depression dataset are encoded as 768-dimensional vectors by pre-trained vision models, and the provided visual descriptors of the D-vlog dataset are 136-dimensional. Let $N$ denotes the sequence length, so each input sample to this submodule is a vector with shape $(N, 49)$ for the E-DAIC dataset, $(N, 768)$ for the Twitter depression dataset and $(N, 136)$ for the D-vlog dataset.

As can be seen from Figure 6.2b, samples are firstly normalized by layer normalization and are encoded by a 1D convlutional layer with 256 filters of kernel size 1 and stride 1. Then, a Gated Linear Unit (GLU) activation is used followed by another 1D convolutional layer with 128 filters of kernel size 3, stride 1 and number of groups 128. Then, a batch normalization and a ReLU activation is used followed by a 1D convlutional layer with 128 filters of kernel size 1, stride 1. Finally, a dropout layer is used and a 4-layer LSTM is adopted to compliment the convlutional layers so both local information and longer-term relations can be obtained for the visual modality.

### 6.1.3 Multi-modal Encoder Module

After obtaining the representations of each modality from the unimodal encoder module, learnt representations are fused by concatenation. For the E-DAIC dataset, textual, audio and visual embeddings are fused into a 1280-dimensional vector for each answer, and for the Twitter depression dataset, textual and visual embeddings are fused into a 1024-dimensional vector for each tweet. Audio and visual features of the D-vlog dataset are constructed as a 512-dimensional vector for each vlog. Then, these multi-modal representations are fed into the multi-modal encoder module to further explore the local and global dependencies, which starts with a transformer submodule encoding the input into attention values as a new representation, followed by a GNN submodule for learning potential new patterns through graph information.

#### Transformer Submodule

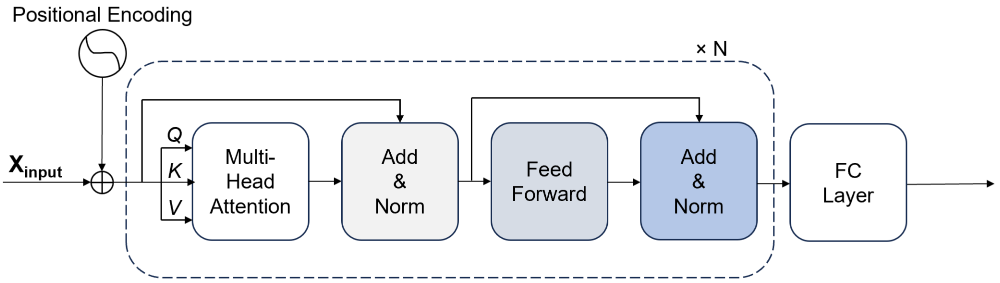


**Figure 6.3:** The transformer submodule.

The transformer submodule is shown in Figure 6.3 which starts with the classic transformer encoder (dash line cell) from the work of [47]. The fused multi-modal representation input $X_{input}$ is sent to the multi-head attention layers after after being added to the positional encoding used for encoding the position of each element of the input sequence. The $Q$, $K$ and $V$ stands for Query, Key, and Value, respectively, obtained by performing linear projections to the multi-modal representation input, which is used for constructing the output embeddings. Let $X$ of shape $(T, d_m)$ as the multi-modal representation input (after addition with positional encoding) of sequence length $T$, the $Q$, $K$ and $V$ are computed using Equation 7.

$$
\begin{aligned}
Q_h &= XW^{h,Q} \\
K_h &= XW^{h,K} \\
V_h &= XW^{h,V}
\end{aligned} \tag{7}
$$

Here, $W^{h,Q}$, $W^{h,K}$ and $W^{h,V}$ are trainable parameters, with $W^{h,Q} \in \mathbb{R}^{d_m \times d_q}$, $W^{h,K} \in \mathbb{R}^{d_m \times d_k}$, $W^{h,V} \in \mathbb{R}^{d_m \times d_v}$, and $d_q = d_k = d_v = d_m$. Then, the output attention matrix of a single head self attention is computed using Equation 8:

$$
head_h = softmax(\frac{Q_h {K_h}^T}{\sqrt{d^k}})V_h \tag{8}
$$

A multi-head attention with $h$ heads is defined in Equation 9, which is simply done by concatenating the attention value obtained from each head and performing another linear operation with a trainable parameter $W^o \in \mathbb{R}^{d_m \times hd_v}$ and $d_q = d_k = d_v = d_m/h$:

$$
MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^o \tag{9}
$$

Then $X$ is added to the obtained multi-head attention value using a residual connection and normalized by layer normalization, and passed to a two-layer feed-forward block with another residual connection followed by another layer normalization. Finally, it is sent to a fully connected layer for downscaling into a 100-dimensional vector.

In our implementation, the transformer encoder is designed to dynamically assign the number of heads with a range from 7 to 15 heads depending on the dimension of $X$. The $N$ presented in Figure 6.3 is set to 4 means that we stack four transformer encoders.

### GNN Submodule

Since the transformer submodule is responsible for modeling global dependencies of the multi-modal representation, a GNN Submodule which is composed of a Relational Graph Convolutional Network (RGCN) [48] and a Graph Transformer [49] is added in charge of modeling local dependencies.

Firstly, we model the output feature vectors obtained from the transformer submodule as a directed graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ so it can be leveraged in the GNN submodule. Each feature vector is considered as a node $v_i \in \mathcal{V}$, with its labeled edges $(v_i, r, v_j) \in \mathcal{E}$ and an edge type $r \in \mathcal{R}$. Two types of edge relations are constructed by taking the past and future 6 utterances or time steps into consideration. Then, nodes along with graphs are sent to the RGCN. The idea behind RGCN is to update the center node feature by taking its neighbor information into consideration in the form of aggregation as the following Equation 10:

$$z'_i = \sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^r} \frac{1}{|\mathcal{N}_i^r|} W_r z_j + W_0 z_i \tag{10}$$

Here $z_i \in \mathbb{R}^{d'_m}$ denotes the node feature obtained from the previous transformer submodule, $W_r$ and $W_0$ are trainable parameters, $\mathcal{N}_i^r$ is the set of neighbors of node $i$ with relation $r$. Then, the updated node features $z'_i$ is sent to a graph transformer, which adopts the multi-head attention of the classic transformer that we explained in the transformer submodule. To update the node feature $z'_i$, the difference is in Equation 7 where $X$ for computing $Q$ is $z'_i$, and $X$ for computing $K$ and $V$ is the neighbor node of $z'_i$. The updated node features are then fed into a batch normalization layer followed by Leaky ReLU activation.

### 6.1.4 Detection Module

The detection module shown in Figure 6.1 contains a LSTM followed by a fully-connected output layer, which takes as input the obtained multi-modal representations from the multi-modal encoder module to estimate the $\hat{Y}$, that is the depression scores or labels.

## 6.2 DPD-E Net

In this section, we introduce the DePressionDetect with EEG Net (DPD-E Net) in detail. Figure 6.4 presents the overall structure of the DPD-E Net, which comprises four individual networks, each making their own predictions, and the results are integrated to yield a final prediction.

### 6.2.1 Model Overview

DPD-E Net is an extended version of the DPD-E Net which is designed to handle EEG, speech and text modalities. In Figure 6.4, $X_{E1}$, $X_{E2}$ and $X_{E3}$ denote three streams of EEG input while $X_A$ and $X_T$ represent the input data from audio and text. $X_{E1}$, $X_{E2}$ are sent to a GAT-based encoder for further encoding, the obtained new embeddings are then passed to the classifier to produce the probability of the subject belonging to the depressed class. $X_{E3}$ is sent to a conv-based encoder followed by another classifier. For $X_A$ and $X_T$, they are processed by the same architecture of DPD-Net using the audio-text modality to produce the depression probabilities. Finally, predictions from each modality $Y_{E1}$, $Y_{E2}$, $Y_{E3}$ and $Y_{AT}$ are fused to generate the decision $\hat{Y}$, with a weighting ratio of 2:2:2:4, respectively. The reason for designing DPD-E Net as an ensemble model using late score fusion is that in the MODMA dataset, EEG data is not collected simultaneously during interviews so the

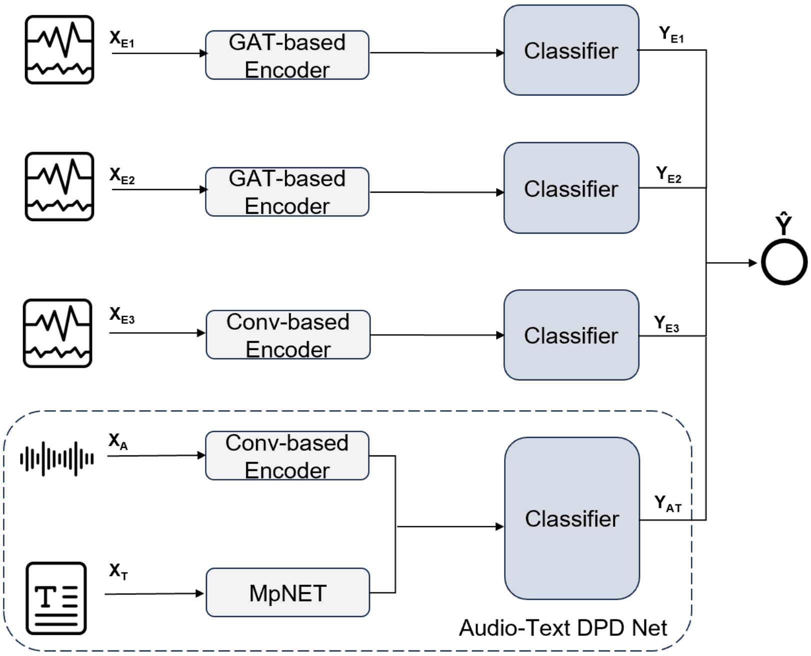


**Figure 6.4:** The proposed DPD-E Net's overall architecture.

audio data and EEG data are not synchronized, which makes it infeasible to early fusing these modalities as employed in DPD Net.

### 6.2.2 EEG Modality

Our DPD-E Net leverages EEG data in a 3-stream fashion. The first stream $X_{E1}$ is a local spatial stream represented by Brain Functional Networks (BFNs) constructed using brain sub-regions right temporal (RT) and left central (LC) while the second stream $X_{E2}$ is a global spatial stream using BFNs constructed from the 16 main electrodes. $X_{E3}$ is the third stream which represents temporal features extracted from EEG signal. The detailed process of constructing Brain Functional Networks (BFNs) and extracting temporal EEG features are presented in Section 5.5.

For EEG data of each subject, channels are considered as nodes and channels' correlation are used for determining edges. Let $C$ denote the number of channels (electrodes) of the brain sub-regions or the number of selected main electrodes, node features of shape $(C, 29)$ along with edge matrix are the input $X_{E1}$ or $X_{E2}$ to the base classifier shown in Figure 6.5. BFNs are firstly passed to the Graph Attention Network (GAT) [50] which updates the node features by aggregating information from first-order neighborhood nodes with different attention coefficients using the following equations:

$$e_{ij} = a^T[Wh_i \parallel Wh_j], j \in \mathcal{N}_i \tag{11}$$

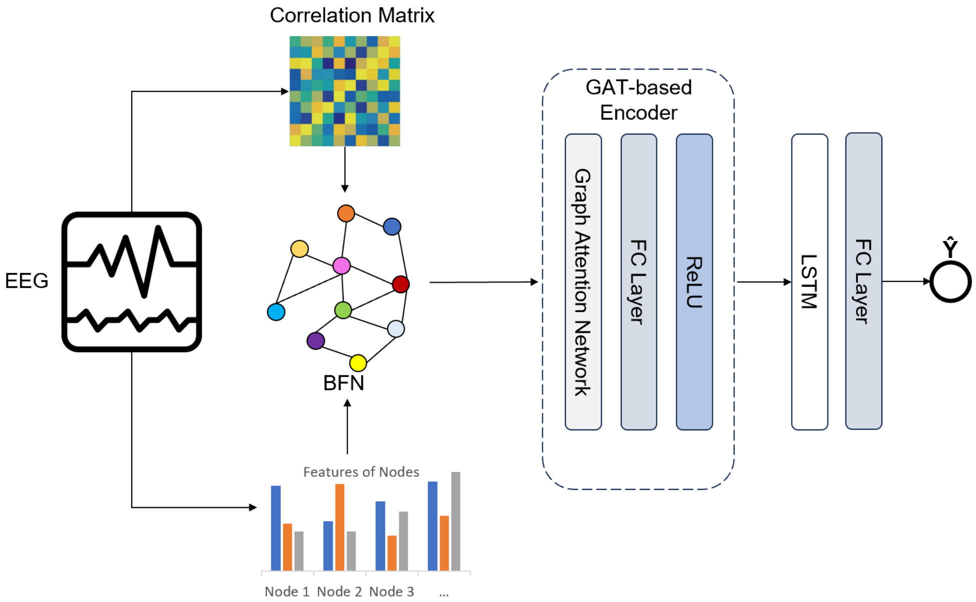


**Figure 6.5:** The EEG-based base classifier of the proposed DPD-E Net for the local spatial stream $X_{E1}$ and the global spatial stream $X_{E2}$.

$$\alpha_{ij} = \frac{exp(LeakyReLU(e_{ij}))}{\sum_{k\in\mathcal{N}_i} exp(LeakyReLU(e_{ik}))} \tag{12}$$

$$h'_i = \overset{K}{\underset{k=1}{||}} \sigma(\sum_{j\in\mathcal{N}_i} \alpha^k_{ij} W^k h_j) \tag{13}$$

Let $i$ be the current node, and $j$ be a node in the set of first-order neighbors $\mathcal{N}_i$. $h_i$ denotes the current node features, and $h_j$ is its first-order neighbor node features. $W$ is a learnable parameter which performs linear transformation to both $h_i$ and $h_j$, and $||$ is the concatenation operation. $a$ is another learnable parameter to map the concatenated vector to real number space. The obtained $e_{ij}$ are then normalized using Equation 12. The final updated node features $h'_i$ are computed using Equation 13 by concatenating each newly obtained node features from $k$ heads. Then, the updated node features $h'_i$ are sent to a fully connected layer followed by a ReLU activation. Finally, DPD Net's detection module consisting of a LSTM and a fully connected layer is used to predict the probability of the subject having depression. In our implementation, 2-head GAT is used for both $X_{E1}$ and $X_{E2}$. During GAT-based encoding, $X_{E1}$ which consists of two types of BFNs constructed from electrodes of two brain sub-regions, are encoded separately by the GAT. Then, the adjusted node features of each sub-region are stacked for the subsequent classification layers.

For the temporal features stream $X_{E3}$, we leverage DPD Net in unimodal way for the classification, but the GNN Submodule is removed since DPD Net employs RGCN in a manner that defines relations by looking back at past utterances and examining future utterances through windows [7], and this approach is not suitable for EEG data. Specifically, temporal features are passed to the Conv-based Submodule designed for audio modality and are further encoded by the Transformer Submodule, and are finally passed to the Detection Module to produce the predicted class.

### 6.2.3 Speech-Text Modality

For audio cues and textual information, we leverage these two modalities in multimodal way by reusing the proposed DPD Net which can be seen from Figure 6.4. Audio and textual cues are firstly encoded by separate unimodal encoders, and then fused into multi-modal embeddings that are further encoded by the multi-modal encoder for the final prediction using the detection module.

# Chapter 7

# Experiment

We conduct extensive experiments to evaluate DPD Net and DPD-E Net. Firstly, we conduct experiments to explore multiple features and the fusion of different obtained features. Also, the best models are compared to several baseline methods. Ablation studies are presented to investigate the contribution of the respective modalities and proposed modules.

## 7.1 Experimental Setup

We conduct experiments on four distinct multi-modal depression datasets. The well-established E-DAIC dataset is used for evaluating the depression detection performance of DPD Net in the clinical setting; the Twitter depression dataset and the D-vlog dataset are used to further evaluate DPD Net's effectiveness in the social media setting; the MODMA dataset is used for validating DPD-E Net's prediction ability.

### Experimental Design

To evaluate and validate our proposed models in a more comprehensive manner, we design four experiments: pre-trained models experiments, modality experiments, comparison to state-of-the-art models and ablation studies of our proposed models. Firstly, we experiment with features obtained from multiple pre-trained models, as the quality of the initial features obtained from these pre-trained models plays an essential role in the depression detection task. In this setting, at least two modalities are used for each dataset, since the main scope of this research is to study multi-modal depression detection. Then, using the best models obtained from the previous experiments, we experiment combinations of each modality to investigate the impact of different modalities on the models' performance, and to verify that different modalities can complement each other in our proposed model. Next, the performance of the proposed models are compared with the state-of-the-art methods. Finally, ablation experiments are conducted to confirm the effectiveness of the modules of DPD

Net and the base classifiers of DPD-E Net.

### Implementation Details

The DPD Net framework is implemented with PyTorch. For the E-DAIC dataset, we use Adam optimizer with cosine annealing with warm restarts scheduler during training. Each model is trained with 1000 epochs using a learning rate of 0.0001, a batch size of 32 and a dropout rate of 0.2. According to the E-DAIC dataset paper, the loss function is designed by leveraging the CCC score, which is simply computed as $1 - CCC_{score}$. For the Twitter depression dataset, Adam optimizer and exponentialLR scheduler are used along with negative log likelihood loss. Each model is trained with 30 epochs using a learning rate of 0.00001, a batch size of 32, a dropout rate of 0.2. For the D-vlog dataset, we use the same settings as we used for the Twitter depression dataset except that each model is trained with 60 epochs. For the MODMA dataset, Adam optimizer and reduceLR scheduler are used during a 20-epochs training with a learning rate of 0.0001, a batch size of 4 and a dropout rate of 0.2.

## 7.2 Experimental Results

### 7.2.1 Pre-trained Models Experiments: Results

The intuition behind the design of these experiments is from the work of [7], where we observe that the textual modality is vital in boosting their model's performance. As our model is based on their work, we suppose that the potential of the textual modality should be explored by experimenting with features obtained from multiple pre-trained language models. Then, we extend this idea to experiment features obtained from different pre-trained vision models to reuse useful patterns exhibited by these pre-trained models. Hence, in general this experiment is designed to investigate what representation methods we should utilize for encoding texts and images.

| | *Modality* | *CCC* | *RMSE* |
|---|---|---|---|
| DistilRoBERTa | T + A | .601 | 5.31 |
| MpNET | T + A | .596 | 5.16 |
| DistilRoBERTa | T + A + V | .617 | 5.51 |
| MpNET | T + A + V | **.682** | 4.79 |

**Table 7.1:** The results on E-DAIC testset using different pre-trained language models and different modality combination.

Experimental results for E-DAIC dataset are presented in Table 7.1. 'A' stands for the audio modality, 'T' is the textual modality and 'V' denotes the visual modality. In the case of textual and audio fusion, DPD Net using features obtained from DistilRoBERTa is able to produce slightly better results compared to the model using MpNET in terms of CCC score. Note that the performance ranking solely relies on CCC score. However, when fusing all the three modalities, the model using MpNET features considerably outperforms its counterpart with the highest CCC score 0.682, and the RMSE is reduced to 4.79. Also, the models using three modalities always achieve better results regardless of the representation methods it leverages. Based on this experiment, we choose DPD Net using MpNET as the best model for further experiments on E-DAIC dataset.

We experiment with multiple pre-trained vision models and language models on the Twitter depression dataset. As shown in Table 7.2, the model using the combination of MpNET as text encoder and CLIP as image encoder outperforms all the other models. The precision obtained using MpNET and FLAVA is reported the same as MpNET-CLIP combination in 0.839 because this is the rounded value, but its real value is actually slightly lower than the MpNET-CLIP model.

With the same visual encoder, models that utilize MpNET achieve better results than models using DistilRoBERTain in terms of F1 score. F1 score is computed by taking both the precision and recall into consideration, also here the experiment used suggests that in general MpNET-based model is the best candidate for this twitter depression detection task. Since the superiority of MpNET can also be observed in the experiments on E-DAIC dataset, we could conclude that MpNET should be chosen over DistilRoBERT as the text encoder of the DPD Net, so the network can work properly on both clinical setting and social media setting. As for the visual encoders, models using CLIP are always be able to produce better F1 score, and for the same reason stated above, we choose CLIP over EfficientNet and FLAVA as the visual encoder to extract initial features from the images.

| *Text Modality* | *Viusal Modality* | *Precision* | *Recall* | *F1* |
|---|---|---|---|---|
| DistilRoBERTa | CLIP | .815 | .810 | .812 |
| DistilRoBERTa | EfficientNet | .796 | .811 | .802 |
| DistilRoBERTa | FLAVA | .794 | .829 | .809 |
| MpNET | CLIP | **.839** | **.874** | **.855** |
| MpNET | EfficientNet | .787 | .836 | .811 |
| MpNET | FLAVA | .839 | .794 | .813 |

**Table 7.2:** The results on Twitter depression dataset using different pre-trained vision and language models.

### 7.2.2 Modality Experiments: Results

The scope of this study is multi-modal depression detection, and we conduct experiments in order to understand the impact of each modality and their combinations on the performance of our proposed models.

Table 7.3 presents the results on the E-DAIC dataset. We can see that for unimodal cases, the model using textual modality outperforms the other two modalities by a significant margin in terms of CCC score and RMSE, which is consistent with the finding that we mentioned at the beginning of the previous paragraph that the textual modality plays an essential role in detecting depression. The model using only the visual modality produces better results compared to the model using only the audio modality. However, after fusing two modalities, the combination of the audio modality with the textual modality gives a better performance than the combination of the visual modality with the textual modality in terms of CCC score. With three modalities, we achieve a CCC score of 0.682 and a RMSE of 4.79 on the E-DAIC dataset, which is the best results obtained from this study on this dataset.

| *Modality* | *CCC* | *RMSE* |
|---|---|---|
| T | .475 | 5.88 |
| A | .161 | 7.42 |
| V | .241 | 6.19 |
| T + A | .596 | 5.16 |
| T + V | .563 | 5.21 |
| T + A + V | **.682** | 4.79 |

**Table 7.3:** The results on E-DAIC testset using unimodal and multi-modal models.

As for results on the Twitter depression dataset shown in Table 7.4, the models using only the textual modality or the visual modality produce comparable results as the textual-based model only slightly outperforms its counterpart in F1 score and recall while the visual-based model is better in terms of precision. DPD Net using both modalities outperforms the unimodal models in all three classification metrics, achieving a performance gain of around 13.5% in terms of F1 score and 23.7% in terms of recall.

From Table 7.5, we can see that on the D-vlog dataset, DPD Net using visual features and audio features outperforms the models using only the audio or visual modality in F1 score, precision and recall.

Table 7.6 lists the experimental results of DPD-E Net on the MODMA dataset. Here 'E1' denotes the global stream using BFN features constructed from the 16 main electrodes, 'E2' denotes the local stream using BFN features constructed from electrodes of prominent

| *Modality* | *Precision* | *Recall* | *F1* |
|---|---|---|---|
| T | .824 | .742 | .777 |
| V | .831 | .721 | .771 |
| T + V | **.834** | **.918** | **.874** |

**Table 7.4:** The results on Twitter depression dataset using unimodal and multi-modal models.

brain sub-regions right temporal (RT), left central (LC), and 'E3' is the temporal stream using EEG features obtained from raw signals. 'A' is audio modality, and 'T' is the textual modality. The MODMA dataset has limited number of subjects and has no train/dev/test split, so experiments on this dataset is conducted using 5-fold cross-validation. Note that the F1 score, precision, and recall are reported as averages over the five folds. According to the results, most of the single-modality models perform poorly in terms of all metrics except for DPD-E Net using textual features. When combining 3-stream EEG features, DPD-E Net achieves a F1 score of around 0.620 and recall of 0.767 but adding audio modality slightly degrades results in these two metrics. With all the listed features, best results in F1 score, precision and recall on the MODMA dataset are produced by DPD-E Net.

| *Modality* | *Precision* | *Recall* | *F1* |
|---|---|---|---|
| V | .681 | .504 | .579 |
| A | .748 | .650 | .696 |
| A + V | **.759** | **.715** | **.736** |

**Table 7.5:** The results on D-vlog dataset using unimodal and multi-modal models.

Based on the experiment results of this section, it is clear that the fusion of different modalities can bring performance gain on all the four datasets, indicating that both DPD Net and DPD-E Net are able to integrating useful information from each modality.

### 7.2.3 Comparison to State-of-the-art Models: Results

After obtaining our best models for each dataset, we compare their performance them with other baseline approaches. The results in Table 7.7 shows that DPD Net with textual and audio modality can achieve around 20.8% higher CCC score compared to the state-of-the-art performance reported by TensorFormer [32], and with three modality fused, the CCC score of DPD Net shows considerable advantages, which is approximately 38.3% higher than the TensorFormer. We rank the performance using CCC score as this metric is used as the only

| *Modality* | $\overline{Precision}$ | $\overline{Recall}$ | $\overline{F1}$ |
|---|---|---|---|
| E1 | .400 | .333 | .313 |
| E2 | .600 | .300 | .393 |
| E3 | .633 | .599 | .567 |
| A | .310 | .467 | .364 |
| T | .853 | .833 | .816 |
| E1 + E2 + E3 | .599 | .767 | .620 |
| E1 + E2 + E3 + A | .633 | .600 | .567 |
| E1 + E2 + E3 + A + T | **.900** | **.900** | **.876** |

**Table 7.6:** The results on MODMA dataset using unimodal and multi-modal models. Here the F1 score, precision, and recall are reported as averages over five folds.

criteria for the depression detection challenge of the E-DAIC dataset paper.

| *Models* | *CCC* | *RMSE* |
|---|---|---|
| TensorFormer (T + A + V) | .493 | 4.31 |
| DPD Net (T + A) | .596 | 5.16 |
| DPD Net (T + A + V) | **.682** | 4.79 |

**Table 7.7:** Comparison to baseline methods on E-DAIC dataset.

The comparison on the Twitter depression dataset is presented in Table 7.8. Since this dataset is released without train-development-test data split, we split the dataset with ratio of 7:1:2 for train set, development set and test set, and re-implement three variants of the baseline method mentioned in Chapter 4 using hyperparameters from the TM Transformer paper. It is obvious that DPD Net outperforms all the other baseline results in terms of F1 score and recall, precision, with a F1 score of 0.874, precision of 0.834 and recall of 0.918.

| *Models* | *Precision* | *Recall* | *F1* |
|---|---|---|---|
| Vanilla TM Transformer | .828 | .904 | .864 |
| Set TM Transformer | .819 | .827 | .823 |
| Time2vec TM Transformer | .824 | .856 | .840 |
| DPD Net | **.834** | **.918** | **.874** |

**Table 7.8:** Comparison to baseline methods on Twitter depression dataset.

According to Table 7.9, DPD Net outperforms its counterpart on the D-vlog dataset in terms of precision but STST transformer [36] achieves higher F1 score and recall according to their published results. The performance of the comparison experiment on these three datasets has proven that DPD Net can achieve sound performance for depression detection on both the clinical setting and the social media setting.

For a fair comparison on the MODMA dataset, we re-implement the baseline method ES Vision Transformer [37] with hyperparameter tuning since the model using their reported hyperparameter values performed poorly on our five folds. The F1 score, precision, and recall are reported as averages over the five folds. As can be seen from Table 7.10, DPD-E Net performs significantly better than the ES Vision Transformer in all the three metrics with a F1 score of 0.876, precision and recall of around 0.900, which demonstrates the potential of incorporating EEG signals with speech data for depression detection.

| *Models* | *Precision* | *Recall* | *F1* |
|---|---|---|---|
| STST Transformer | .725 | **.776** | **.750** |
| DPD Net | **.759** | .715 | .736 |

**Table 7.9:** Comparison to baseline methods on D-vlog dataset.

| *Models* | $\overline{\textit{Precision}}$ | $\overline{\textit{Recall}}$ | $\overline{\textit{F1}}$ |
|---|---|---|---|
| ES Vision Transformer | .566 | .800 | .651 |
| DPD-E Net | **.900** | **.900** | **.876** |

**Table 7.10:** Comparison to baseline methods on MODMA dataset. Here the F1 score, precision, and recall are reported as averages over five folds.

### 7.2.4 Ablation Study: Contribution of Modules Experiments

For DPD Net, we conduct an ablation study to investigate the contribution of each of the sub-modules. Experiments are only conducted on the E-DAIC dataset as it is currently the most used benchmark depression dataset.

Firstly, we explore the effect of the Conv-based Submodule which is the core component of DPD Net's unimodal encoder module. The Conv-based Submodule consists of CNN blocks and LSTM, and to study their effectiveness, we remove each of them or omit both of them. Table 7.11 presents the results after changing these blocks. It is obvious that only using the Conv block or LSTM block, DPD Net performs to a certain extent with a CCC score around to 0.4. However, after removing both the Conv block and LSTM block, the model performs poorly with the CCC score dropping from 0.682 to 0.141, which suggests that the proposed unimodal encoder module is able to significantly improve the efficiency of DPD Net and its inclusion is important to the DPD Net's performance.

| Conv-based Submodule | | | |
|---|---|---|---|
| *Conv Block* | *LSTM* | *CCC* | *RMSE* |
| - | - | .141 | 7.72 |
| - | ✓ | .441 | 6.22 |
| ✓ | - | .453 | 5.91 |
| ✓ | ✓ | .682 | 4.79 |

**Table 7.11:** Ablation study on Conv-based Submodule of DPD Net on the E-DAIC dataset.

Next, we investigate the impact of the Transformer Submodule and GNN Submodule by removing one of them separately or removing them both while keeping other components of DPD Net unchanged. Results shown in Table 7.12 reveal that both the Transformer Submodule and GNN Submodule play an essential role in multi-modal encoding as deleting one of them leads to low CCC scores of 0.106 and 0.206, respectively. Also, it is worth noting that using only one of submodules is almost equivalent to removing the whole multi-modal encoder module, which means the information obtained from the Transformer Submodule and GNN Submodule can complement each other to produce more informative embeddings for depression detection.

| *Transformer Submodule* | *GNN Submodule* | *CCC* | *RMSE* |
|---|---|---|---|
| - | - | .160 | 7.59 |
| - | ✓ | .106 | 7.61 |
| ✓ | - | .206 | 7.36 |
| ✓ | ✓ | .682 | 4.79 |

**Table 7.12:** Ablation study on Transformer Submodule and GNN Submodule of DPD Net on the E-DAIC dataset.

As for the ensemble model DPD-E Net, we study the effects of the base classifiers on the MODMA dataset. In Table 7.13, the last row presents the performance of the complete DPD-E Net and the first four rows are the results of removing each of the base classifiers. Note that again all the metrics are reported as averages over the five folds. The F1 score after deleting the classifier based on local spatial stream drops from 0.876 to 0.748. Thus including this base classifier could bring a performance gain of around 15%. The deletion of the global spatial stream classifier or the EEG temporal features stream classifier leads to an overall decreased in performance, with F1 scores decreasing around 7%. These results suggest that EEG-based base classifiers improve DPD-E Net's performance. Further, we find that DPD-E Net without the Speech-Text stream classifier gives the most significant decline in performance, with a F1 score, precision, recall decreasing from 0.876 to 0.620, from 0.900 to 0.767, from 0.900 to 0.599, respectively. This approximately 29% decline in F1 score indicates that the Speech-Text stream classifier is the most important classifier, which also confirms the finding that we mentioned above that the involvement of textual features plays an essential role in detecting depression.

| *Classifiers* | $\overline{\textit{Precision}}$ | $\overline{\textit{Recall}}$ | $\overline{\textit{F1}}$ |
|---|---|---|---|
| - Local spatial stream | .850 | .700 | .748 |
| - Global spatial stream | .870 | .833 | .815 |
| - EEG temporal features stream | .920 | .800 | .817 |
| - Speech-Text stream | .599 | .767 | .620 |
| DPD-E Net | .900 | .900 | .876 |

**Table 7.13:** Ablation study on the base classifiers of DPD-E Net. Here the F1 score, precision, and recall are reported as averages over five folds.

# Chapter 8

# Conclusion

In this work, a GNN-enhanced Transformer model named DPD-Net is proposed to solve the problem of automatic depression detection using multi-modal data. The key idea of DPD-Net is to use the unimodal encoder module for encoding each single modality and then use the multi-modal encoder module to further extract useful information from multi-modal features using a transformer sub-module that captures long-term relations and a GNN sub-module that retains local dependencies. This is the first attempt to our knowledge to employ GNNs along with transformers for automatic depression detection. Also, unlike previous works which mostly are tested on a single dataset, DPD Net can work across different datasets under two different application settings, including the clinical setting and the social media setting. Furthermore, its extended version DPD-E Net can be applied to an additional EEG modality. We design two models so the proposed network can be suitable for different real-world scenarios and might have the potential to be deployed in practical use cases.

Our ablation studies demonstrate the advantages of the proposed sub-modules and base classifiers, and the effectiveness of combining diverse data modalities for automatic depression detection. Comparisons to other baseline methods show that DPD Net and DPD-E Net can outperform the state-of-the-art models on three datasets: E-DAIC dataset, Twitter depression dataset and MODMA dataset, and achieve competitive performance on the D-vlog dataset.

Based on the modality experimental results, we observe that even though the involvement of audio, visual and EEG modality can improve the prediction results, the performance of both DPD Net and DPD-E Net highly rely on the the text modality. DPD Net only produces comparable results but is not able to beat the baseline method in all metrics on the D-vlog dataset, which is the only dataset in this work where the textual modality is not included. In the future, we hope to combine textual data to improve the performance on the D-vlog dataset and add more explainability to the proposed models.

# Bibliography

[1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLOS Medicine*, vol. 3, pp. 1–20, 11 2006.

[2] A. Kleiboer, J. Smit, J. Bosmans, J. Ruwaard, and G. Andersson, "European comparative effectiveness research on blended depression treatment versus treatment-as-usual (e-compared): study protocol for a randomized controlled, non-inferiority trial in eight european countries," *Trials*, vol. 17, pp. 1–10, Aug. 2016.

[3] D. Santomauro, A. Mantilla Herrera, J. Shadid, and P. Zheng, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic," *The Lancet*, vol. 398, 10 2021.

[4] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, (New York, NY, USA), p. 3–12, Association for Computing Machinery, 2019.

[5] K. Smith, P. Renshaw, and J. Bilello, "The diagnosis of depression: Current and emerging methods," *Comprehensive psychiatry*, vol. 54, 08 2012.

[6] H. Sun, J. Liu, S. Chai, Z. Qiu, L. Lin, X. Huang, and Y. Chen, "Multi-modal adaptive fusion transformer network for the estimation of depression level," *Sensors (Basel, Switzerland)*, vol. 21, 2021.

[7] A. Joshi, A. Bhat, A. Jain, A. Singh, and A. Modi, "COGMEN: COntextualized GNN based multimodal emotion recognitioN," (Seattle, United States), pp. 4148–4164, Association for Computational Linguistics, July 2022.

[8] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy*, vol. 22, no. 6, 2020.

[9] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, (Reykjavik, Iceland), pp. 3123–3128, European Language Resources Association (ELRA), May 2014.

[10] Z. Zhao, Z. Bao, Z. Zhang, J. Deng, N. Cummins, H. Wang, J. Tao, and B. Schuller, "Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 423–434, 2020.

[11] F. Yin, J. Du, X. Xu, and L. Zhao, "Depression detection in speech using transformer and parallel convolutional neural networks," *Electronics*, vol. 12, no. 2, 2023.

[12] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 542–552, 2020.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, 2016.

[14] Y. Wang, J. Ma, B. Hao, P. Hu, X. Wang, J. Mei, and S. Li, "Automatic depression detection via facial expressions using multiple instance learning," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1933–1936, 2020.

[15] W. Carneiro de Melo, E. Granger, and M. B. Lopez, "Encoding temporal information for automatic depression recognition from facial analysis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1080–1084, 2020.

[16] C. Casado, M. L. Cañellas, and M. B. López, "Depression recognition using remote photoplethysmography from facial videos," *IEEE Transactions on Affective Computing*, pp. 1–13, 2023.

[17] L. Ansari, S. Ji, Q. Chen, and E. Cambria, "Ensemble hybrid learning methods for automated depression detection," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 1, pp. 211–219, 2023.

[18] Y. Zhang, Y. He, L. Rong, and Y. Ding, "A hybrid model for depression detection with transformer and bi-directional long short-term memory," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2727–2734, 2022.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.

[20] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, p. 2, 2019.

[21] E. Toto, M. Tlachac, and E. A. Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, (New York, NY, USA), Association for Computing Machinery, 2021.

[22] S. Senn, M. Tlachac, R. Flores, and E. Rundensteiner, "Ensembles of bert for depression classification," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 4691–4694, 2022.

[23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

[24] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, (New York, NY, USA), p. 69–76, Association for Computing Machinery, 2017.

[25] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, (New York, NY, USA), p. 55–63, Association for Computing Machinery, 2019.

[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[27] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE spoken language technology workshop (SLT)*, pp. 1021–1028, IEEE, 2018.

[28] F. Zhao, H. Pan, N. Li, X. Chen, H. Zhang, N. Mao, and Y. Ren, "High-order brain functional network for electroencephalography-based diagnosis of major depressive disorder," *Frontiers in Neuroscience*, vol. 16, 2022.

[29] H. Cai, Z. Yuan, Y. Gao, S. Shuting, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, Q. Zhao, Z. Liu, Z. Yao, M. Yang, H. Peng, Z. Jing, X. Zhang, G. Gao, F. Zheng,

and B. Hu, "A multi-modal open dataset for mental-disorder analysis," *Scientific Data*, vol. 9, p. 178, 04 2022.

[30] J. Yoon, C. Kang, S. Kim, and J. Han, "D-vlog: Multimodal vlog dataset for depression detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 12226–12234, 2022.

[31] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in twitter," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 110–117, 2019.

[32] H. Sun, Y.-W. Chen, and L. Lin, "Tensorformer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection," *IEEE Transactions on Affective Computing*, pp. 1–11, 2022.

[33] A.-M. Bucur, A. Cosma, P. Rosso, and L. P. Dinu, "It's just a matter of time: Detecting depression with time-enriched multimodal transformers," in *European Conference on Information Retrieval*, pp. 200–215, Springer, 2023.

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[35] R. Kamath, A. Ghoshal, S. Eswaran, and P. B. Honnavalli, "Emoroberta: An enhanced emotion detection model using roberta," in *IEEE International Conference on Electronics, Computing and Communication Technologies*, 2022.

[36] Y. Tao, M. Yang, Y. Wu, K. Lee, A. Kline, and B. Hu, "Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer," *Digital Communications and Networks*, 2023.

[37] A. Qayyum, I. Razzak, M. Tanveer, M. Mazher, and B. Alhaqbani, "High-density electroencephalography and speech signal based deep framework for clinical depression diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 4, pp. 2587–2597, 2023.

[38] B. Zhang, D. Wei, G. Yan, X. Li, Y. Su, and H. Cai, "Spatial–temporal eeg fusion based on neural network for major depressive disorder detection," *Interdisciplinary Sciences: Computational Life Sciences*, pp. 1–18, 2023.

[39] S. Mahato and S. Paul, "Detection of major depressive disorder using linear and non-linear features from eeg signals," *Microsystem Technologies*, vol. 25, pp. 1065–1076, 2019.

[40] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, p. 103107, 2022.

[41] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[42] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[43] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857–16867, 2020.

[44] S. Sun, J. Li, H. Chen, T. Gong, X. Li, and B. Hu, "A study of resting-state eeg biomarkers for depression recognition," *ArXiv*, vol. abs/2002.11039, 2020.

[45] B. Zhang, H. Cai, Y. Song, L. Tao, and Y. Li, "Computer-aided recognition based on decision-level multimodal fusion for depression," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3466–3477, 2022.

[46] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," pp. 5036–5040, 10 2020.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[48] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *The Semantic Web: 15th International Conference*, pp. 593–607, Springer, 2018.

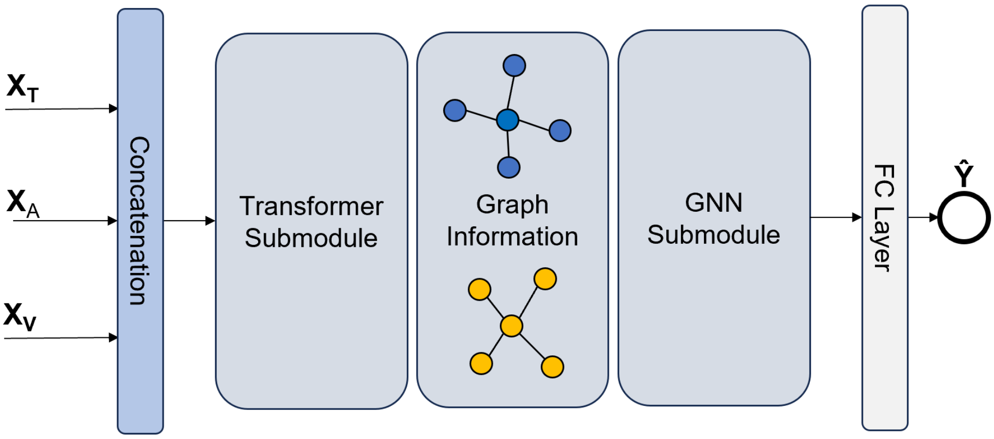
[49] Y. Shi, Z. Huang, W. Wang, H. Zhong, S. Feng, and Y. Sun, "Masked label prediction: Unified massage passing model for semi-supervised classification," *ArXiv*, vol. abs/2009.03509, 2020.

[50] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018.

# Appendix A

# COGMEN Architecture

The multi-modal encoder module of DPD Net is based on the COGMEN model [7]. Here we present the architecture of this model.



The architecture of the COGMEN.

# Appendix B

# Data Pre-processing of the Baselines

Baseline methods pre-process the datasets in different ways. For the E-DAIC dataset which provides official train/dev/test split, Tensorformer pre-process the data by only keeping the first 100 words of each utterance and its corresponding audio-visual features [32]. For the D-vlog dataset which also provides official data splits, STST Transformer discard the samples which have all-zero values [36]. As for the Twitter depression dataset and the MODMA dataset which have no official divisions of the data, since we reproduced TM Transformer and ES Vision Transformer, and conduct the experiments using our own pre-processed datasets, the pre-process procedures of these two models are not presented here.

# Appendix C

# Reported Results of the Baselines

Note that in Section 7.2.3, for the comparison with Tensorformer on the E-DAIC dataset and the comparison with STST Transofermer on the D-vlog dataset, we present their reported results from the original papers. For the comparison with TM Transformer on the Twitter depression dataset and the comparison with ES Vision Transofermer on the MODMA dataset, we only present experimental results from our implementations of these two baselines. Here, the results published in the original papers using their own data splits are shown in the following tables.

| *Models* | *Precision* | *Recall* | *F1* |
|---|---|---|---|
| Vanilla TM Transformer | .868 | .905 | .886 |
| Set TM Transformer | .921 | .934 | .927 |
| Time2vec TM Transformer | .931 | .931 | .931 |

The reported results of TM Transformer on the Twitter depression dataset.

| *Models* | *Precision* | *Recall* | *F1* |
|---|---|---|---|
| ES Vision Transformer | .977 | .973 | .973 |

The reported results of ES Vision Transformer on the MODMA dataset.