



Multi-modal Depression Estimation Based on Sub-attentional Fusion

Ping-Cheng Wei^(✉), Kunyu Peng, Alina Roitberg, Kailun Yang,
Jiaming Zhang, and Rainer Stiefelhagen

Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology,
Karlsruhe, Germany

pingcheng.wei99@gmail.com, {kunyu.peng,alina.roitberg,kailun.yang,jiaming.zhang,rainer.stiefelhagen}@kit.edu

<https://github.com/PingCheng-Wei/DepressionEstimation>

Abstract. Failure to timely diagnose and effectively treat depression leads to over 280 million people suffering from this psychological disorder worldwide. The information cues of depression can be harvested from diverse heterogeneous resources, *e.g.*, audio, visual, and textual data, raising demand for new effective multi-modal fusion approaches for automatic estimation. In this work, we tackle the task of automatically identifying depression from multi-modal data and introduce a sub-attention mechanism for linking heterogeneous information while leveraging Convolutional Bidirectional LSTM as our backbone. To validate this idea, we conduct extensive experiments on the public DAIC-WOZ benchmark for depression assessment featuring different evaluation modes and taking gender-specific biases into account. The proposed model yields effective results with 0.89 precision and 0.70 F1-score in detecting major depression and 4.92 MAE in estimating the severity. Our attention-based fusion module consistently outperforms conventional late fusion approaches and achieves competitive performance compared to the previously published depression estimation frameworks, while learning to diagnose the disorder end-to-end and relying on far fewer preprocessing steps.

Keywords: Depression estimation · Multi-modal fusion · ConvBiLSTM · Speech recognition · Computer vision · Natural language processing

1 Introduction

Depression is a common and serious medical condition, negatively impacting the daily lives of >280 million people according to the World Health Organization (WHO) [46]. The severe manifestation of depression is referred to as Major Depressive Disorder (MDD) or Major Depression (MD), which is defined as a

P.-C. Wei and K. Peng—The first two authors contribute equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-25075-0_42.

mental state of pervasive and persistent low mood, accompanied by the possibility of aversion to activity [4,30].

MD is hard to diagnose: the common symptoms, *e.g.*, pessimism, low self-esteem, and cynical behaviour, are more subjective and therefore more difficult to detect compared to most physical illnesses. As a consequence, around 33% of patients with depression are not recognized during clinical diagnostic procedures and less than 40% of people with this condition receive proper treatment [23]. Fortunately, depression is treatable under certain conditions such as early diagnosis [19,27], but making such large-scale diagnostics accessible for the majority of the population will greatly increase the work pressure of psychologists, demonstrating the importance of assistive tools for end-to-end automated estimation of MD. The release of the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset [10] enabled systematic development and evaluation of learning-based approaches for depression assessment. In this dataset, depression cues are learned from audio, visual and textual data, showing promising performance for MD estimation, while audio has been the most common modality in the past work (*e.g.*, DepAudioNet [29]). Different cues of MD can be harvested from different types of data: for example, depressed patients tend to speak monotonously, feebly, or anxiously and show less head motion, eye contact, or smiling [15,45]. Given the complementary nature of different data sources, multimodality is of key importance for improving automatic depression assessment models, but *how to fuse the information* becomes an important research direction and is the main motivation of our work.

In this work, we introduce a neural network-based multi-modal architecture for MD estimation. We start by choosing DepAudioNet [29], one of the best models in the AVEC 2016 challenge [43], as our research baseline for MD estimation and propose a novel mechanism for fusing visual, audio, and textual data via attention-based building blocks. The proposed framework comprises a Convolutional Bidirectional LSTM (ConvBiLSTM) as our feature extraction backbone and sub-attentional fusion leveraging attention for each individual MD sub-score estimation head. We conduct extensive experiments on the public DAIC-WOZ dataset [10], comparing the proposed sub-attention-based fusion strategy with different score-based late fusion techniques (*e.g.*, addition, concatenation). We study gender bias in MD estimation and compare our model to previously published approaches, which, in contrast to our work, rely on far stronger feature engineering, preprocessing and data cleaning steps. We further validate our model separately for participant-level and clip-level MD estimation. Our proposed sub-attentional fusion model outperforms other data fusion techniques and yields competitive performance for MD estimation, while featuring far less feature engineering and data preprocessing than previously published approaches, for the purpose of a more accessible end-to-end automated depression assessment.

2 Related Work

Benefiting from the rapid development of data-driven methods, depression estimation through learning-based techniques has attracted remarkable attention.

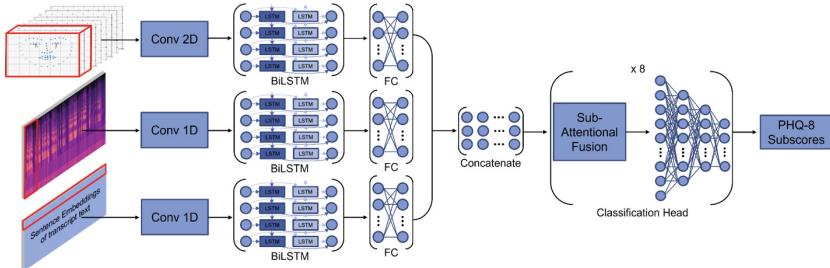


Fig. 1. An illustration of the proposed Sub-attentional ConvBiLSTM model structure. It has inputs from three different modalities and outputs PHQ-8 Subscores corresponding to the severity of 8 major depression symptoms. For the backbone, a serial combination of CNN and BiLSTM layers is exploited. After the fully-connected (FC) layers, the extracted features are then concatenated and processed by 8 different attentional fusion layers, followed by 8 individual classifiers. This hierarchical structure maximizes the effectiveness of multi-modal features with global- and local attentional fusion.

The recently proposed methods targeting depression estimation can be divided into four groups: text-based, audio-based, vision-based, and multi-modal methods, enabled by the release of the multi-modal depression estimation dataset DAIC-WOZ [10]. According to [39], even slight differences of the psychological state have the potential to cause a noticeable change in the acoustic domain, which makes audio data a competitive modality to be used as input to depression estimation frameworks [7, 8, 13, 25, 29, 39, 48]. Mel-cepstral features and formant-frequency tracks are two main arousal representations introduced by Williamson *et al.* [48] harvesting discriminative arousal cues from vocal tract resonant frequencies and spectral dynamics. Scherer *et al.* [39] introduced four voice-based features for the psychological distress, which are subsequently classified by Support Vector Machine (SVM). Since deep learning has gradually took over the pattern recognition field in recent years, remarkable progress has been achieved with the help of deep neural networks instead of hand-crafted feature-based approaches, with an overview of such methods based on audio data provided by He *et al.* [22]. Excellent performance was reported by Ma *et al.* [29] who proposed DepAudioNet, an end-to-end depression estimation model using Convolutional Neural Network (CNN) and Long Short-Term Memory network (LSTM). Saidi *et al.* [36] aim at the depression degree estimation fusion via multivariate regression. Sardari *et al.* [38] proposed an audio-based depression detection architecture using a convolutional autoencoder. Apart from audio, text data is another popular source of depression cues. A text-based multi-task Bidirectional Gate Recurrent Unit (BGRU) network for depression estimation is proposed by Dinkel *et al.* [11]. Salimath *et al.* [37] proposed a metric to quantify the depression severity by utilizing negative sentences. Visual cues, which are mainly extracted from facial key points [5, 18, 20, 34, 50] or raw video data [1, 3, 31, 32], also serve for depression estimation by capturing slight facial expression changes. Facial Action Units (FAUs), facial landmarks, head pose and gaze direction are

utilized as the CNN input for visual data based approach [12]. Xie *et al.* [49] leveraged video data to interpret depression from question-wise long-term video recordings using 3D CNNs. In order to combine cues from different modality types, several fusion architectures are also proposed by the researchers in recent years [2, 16, 24, 35, 40, 44]. A deep multi-modal network for MD assessment is introduced by [42]. He *et al.* [21] realized a multi-modal depression estimation by combining visual and audio cues. A deeper causal neural network is proposed by Gong *et al.* [16] for the fusion of different modalities. Drawing inspiration from recent progress in conventional image classification [9], we tackle the depression estimation task via cross-modality attention-based fusion. Furthermore, most of the existing single-modal and multi-modal approaches rely on heavy data preprocessing techniques which violate the end-to-end principle. In contrast to these approaches, we tackled MD assessment by directly using the nearly raw dataset with less data cleaning techniques to train and test the performance of the investigated baselines and our proposed model.

3 Model

In this paper, we present a novel deep learning model named Sub-attentional ConvBiLSTM - an attentional multi-modal architecture featuring late fusion of extracted representations from three different modalities. An overview of our model is provided in Fig. 1. Benefiting from its hierarchical structure, Sub-attentional ConvBiLSTM is highly effective in depression estimation while keeping a low computational load, which will be unfolded in Sec. 4. In addition, two techniques for increasing the performance are introduced, *i.e.*, “Multi-path Uncertainty-aware Score Distributions Learning (MUSDL) [41]” and “Sharpness-Aware Minimization (SAM) [14]”. MUSDL is a specific score distribution generation technique for converting each hard-label in the Ground Truth (GT) to a soft-label score distribution for soft decisions. As for SAM, it is a second-order optimization method, which is specifically devised and has been proven [6] to improve the generalization ability of the model, even just training on a small dataset (which is a common case in depression estimation).

3.1 Sub-attentional ConvBiLSTM

For the input of Sub-attentional ConvBiLSTM, three different feature domains, *i.e.*, log-mel spectrograms (audio), micro-facial expressions (visual), and sentence embeddings (text), have been deployed. These inputs will then be processed by each backbone to extract the higher-level representation of each feature. The backbone chosen here is based on “DepAudioNet” from Ma *et al.* [29], which was one of the best models in the AVEC 2016 [43] by exploiting a deep-learning-based approach with solely acoustic features for depression detection. In general, DepAudioNet uses a serial combination of 1D-CNN layers and LSTM layers with a 2-dimensional input format. We have further improved the model by transforming

LSTM layers into Bidirectional LSTM (BiLSTM) layers and enabling the applicability to a 3D input format for visual input by utilizing 2D-CNN layers.

For the audio and text branch, 1D-CNN layers are first used to provide translation-equivariant responses of a low-level feature map, whose kernel size k is 3, indicating that several short-term features are captured at these layers. The visual branch, however, uses a 2D-CNN layers with a kernel size $k = (72 \times 3)$, as all 72 key points are perceived as a whole and the kernel slides through the visual data solely along the time-axis, focusing on extracting temporal changes between each frame to provide local attention. Then, batch normalization is performed to regularize the intermediate representation of the features to a standard normal distribution, followed by a nonlinear transformation with the Rectified Linear Unit (ReLU), an activation function defined as $f(x) = \max(0, x)$. To further reduce the dimensionality, a max-pooling layer is applied to down-sample the input representation of the feature map. A BiLSTM layer together with an FC layer is stacked at the end of the backbone structure, for an objective of harvesting long-range variability in each modality along the time-axis and retrieving effective features from each branch. After all features from each modality are extracted, they are then concatenated in parallel to form a feature map as input to the subsequent late fusion layer.

Considering the superiority of weighting fusion methods over the traditional fusion methods [28] and for more effective deployment of such feature map of multi-modal data, we insert attentional fusion layers inspired by Dai *et al.* [9] to realize attentional information interaction between each modality. Details of such a structure as well as its functionality, will be discussed in Sec. 3.2. In this Sub-Attentional ConvBiLSTM model, 8 different attentional fusion layers are exploited, connected with 8 different output heads, respectively, which correspond to the subclass number of the PHQ-8 Subscores. With this structure, each sub-attentional fusion block will be trained to have the competence in focusing on distinct depression cues from different modalities. For output heads, classifiers are used to predict the PHQ-8 Subscores and the final PHQ-8 Score, the indicator of the severity of depression, as well as the final PHQ-8 Binary, the binary state of having MD, are further derived based on their definitions [26].

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

3.2 Attentional Fusion Layer

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

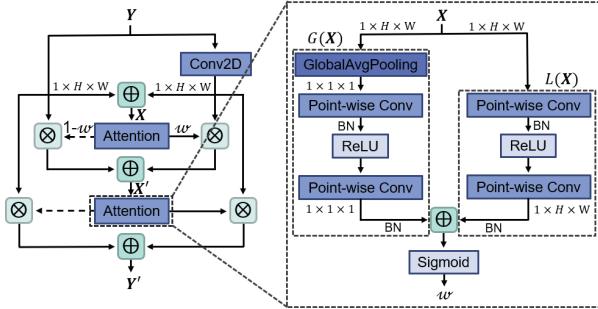


Fig. 2. An illustration of the proposed attentional fusion block. A hierarchical structure is built from top to bottom, splitting into three different layers, which is inspired by the work from Dai *et al.* [9].

the second layer. C denotes number of channel, which is 1 in our case, and $H \times W$ denotes the size of the feature map.

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

$$w = \sigma(G(\mathbf{X}) \oplus L(\mathbf{X})), \quad (1)$$

which is an aggregation of the global feature attention $G(\mathbf{X}) \in \mathbb{R}^C$ and the local channel attention $L(\mathbf{X}) \in \mathbb{R}^{C \times H \times W}$ transformed through a sigmoid activation function σ as shown in the magnified illustration in the Fig. 2. The global feature attention can be obtained via the following equation:

$$G(\mathbf{X}) = BN(W_2 \cdot \text{ReLU}(BN(W_1 \cdot g(\mathbf{X})))), \quad (2)$$

As the name implies, the global feature context of the intermediate feature map will be first extracted through a Global Average Pooling (GAP) block $g(\mathbf{X}) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \mathbf{X}_{[:,i,j]}$, followed by dimension decreasing and increasing blocks, *i.e.*, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, with a Rectified Linear Unit (ReLU) layer in the middle. r is the channel reduction ratio and both dimension decreasing and increasing blocks are in fact implemented as a point-wise convolution (*PWConv*). After each block, batch normalization (*BN*) is applied. As for the local feature attention, a similar structure can be established via excluding the GAP block $g(\mathbf{X})$. Hence, the function could be summarized as:

$$L(\mathbf{X}) = BN(PWConv_2 \cdot \text{ReLU}(BN(PWConv_1 \cdot \mathbf{X}))). \quad (3)$$

Moreover, it is noteworthy that the resulting local attentional weight of $L(\mathbf{X})$ has the identical shape ($C \times H \times W$) as the input, which can be trained to preserve and highlight the subtle details of depression cues from the intermediate feature map. After the channel attention weight w is derived, the complementary channel attention weight ($1 - w$) is also calculated, as denoted with the dashed line in

Fig. 2. The refined feature (RF) as well as the complementary refined feature (RF^c) can then be calculated via the following equations:

$$\begin{aligned} RF &= Conv(\mathbf{Y}) \otimes w = Conv(\mathbf{Y}) \otimes \sigma(G(\mathbf{X}) \oplus L(\mathbf{X})), \\ RF^c &= \mathbf{Y} \otimes (1 - w) = \mathbf{Y} \otimes (1 - \sigma(G(\mathbf{X}) \oplus L(\mathbf{X}))). \end{aligned} \quad (4)$$

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

$$\mathbf{X}' = RF \oplus RF^c = Conv(\mathbf{Y}) \otimes w \oplus \mathbf{Y} \otimes (1 - w). \quad (5)$$

In the third layer, the attentional process explained previously will be performed again to further improve and accentuate depressive characteristics in the transitional attentional feature \mathbf{X}' of multimodality. Therefore, the ultimate attentional feature fusion output $\mathbf{Y}' \in \mathbb{R}^{C \times H \times W}$ can be expressed as:

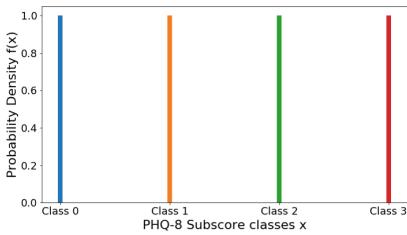
$$\mathbf{Y}' = Conv(\mathbf{Y}') \otimes w' \oplus \mathbf{Y}' \otimes (1 - w'), \quad (6)$$

where w' is the channel attention weight outputted from \mathbf{X}' . Thereby, an adaptive multi-modal interaction is realized, which is beneficial for accurate depression estimation. In our model, the attentional feature fusion output \mathbf{Y}' will then be input to the 8 classification heads for the PHQ-8 Subscores classification.

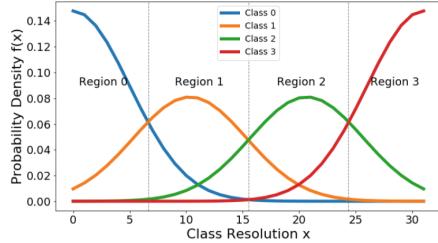
3.3 MUSDL

MUSDL stands for Multi-path Uncertainty-aware Score Distributions Learning proposed by Tang *et al.* [41]. This method converts the hard-label score in GT to a soft-label distribution for soft decisions and has demonstrated effectiveness to solve the intrinsic ambiguity in GT and boost the performance. We flexibly adopt it to reinforce our depression estimation model in harvesting discriminative cues.

Given a classification GT $\mathbf{s}_{GT} \in \mathbb{N}^n$ of an interview clip containing a set of n hard-label scores $s \in \mathbb{N}$: $\mathbf{s}_{GT} = [s_1, s_2, \dots, s_n]$, each score s in the GT will be transformed into a Gaussian-distribution-like soft-label vector $\vec{s} \in \mathbb{R}^{m'}$. It follows $\vec{s} = \mathcal{N}(\mu, \sigma^2)$ with a mean $\mu = s$ and a standard deviation of σ , where each hard-label $\{s \in \mathbb{N} | 0 \leq s < m\}$ is an integer and the soft-label $\vec{s} = [s'_1, s'_2, \dots, s'_{m'}]$ is a discrete set of scores with $\{s' \in \mathbb{R} | 0 \leq s' \leq 1\}$. Here, σ is a hyper-parameter which serves as the level of uncertainty for assessing a clip and $m, m' \in \mathbb{N}$ denote the class resolution or the number of the classes before and after the soft-label transformation. The transformed ratio $r \in \mathbb{R}$ can be derived via $r = \frac{m'}{m}$, which should be equal or greater than 1, indicating an unchanging or expansion of class resolution. The higher the ratio is, the smoother the distribution curve becomes, leading to a better soft-decision strategy performance. In the end, by uniformly discretizing each hard label s in \mathbf{s}_{GT} into a normalized soft-label vector \vec{s} , a matrix of n Gaussian distributions $\mathbf{S}_{GT}^{soft} \in \mathbb{R}^{n \times m'}$ can be obtained. The overall transformation process can be summarized and expressed via:



(a) GT hard-label.



(b) Transformed GT soft-label.

Fig. 3. An overview of label transformation through MUSDL. (a) is the 4 hard-label subclasses in PHQ-8 and (b) is the converted soft-label for each subclass.

$$\mathbf{s}_{GT} = [s_1, s_2, \dots, s_n]$$

⊕
Label Transformation

$$\begin{aligned} \mathbf{S}_{GT}^{soft} &= [\vec{s}_1, \vec{s}_2, \dots, \vec{s}_n] \\ &= \begin{bmatrix} s'_{1,1} & s'_{1,2} & \cdots & s'_{1,m'} \\ s'_{2,1} & s'_{2,2} & \cdots & s'_{2,m'} \\ \vdots & \vdots & \ddots & \vdots \\ s'_{n,1} & s'_{n,2} & \cdots & s'_{n,m'} \end{bmatrix}. \end{aligned} \quad (7)$$

In this work, n is equal to 8 and m is 4 (class 0 to class 3) in accordance with the definition of PHQ-8 Subscores. The standard deviation σ is set to 5 and the transformed ratio r is 8, indicating that the number of the class is expended from $m = 4$ to $m' = 32$. The final transformed label is illustrated in Fig. 3. One can notice that before the transformation, the hard-label GT of 4 different classes is given. After the transformation, a probability density function of the normal distribution is generated. Furthermore, during the training stage, all of the 8 different classification heads are trained to predict the probability between the 4 different depressive classes of the corresponding subscore with the softmax-function: $\mathbf{S}_{pred}^{soft} = [\vec{s}_{1,pred}, \vec{s}_{2,pred}, \dots, \vec{s}_{n,pred}]$. The learning loss is then calculated through pointwise KL divergence between \mathbf{S}_{GT}^{soft} and \mathbf{S}_{pred}^{soft} , which can be computed as:

$$KL(\mathbf{S}_{GT}^{soft} \parallel \mathbf{S}_{pred}^{soft}) = \sum_{i=1}^n \sum_{j=1}^{m'} \vec{s}_{i,j} \cdot \log \frac{\vec{s}_{i,j}}{\vec{s}_{i,j,pred}}. \quad (8)$$

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

$$\mathbf{s}_{pred} = \lfloor \arg \max_{\vec{s}_{i,pred}} \{\vec{s}_{1,pred}, \vec{s}_{2,pred}, \dots, \vec{s}_{n,pred}\} / r \rfloor. \quad (9)$$

3.4 SAM

The problem with the first-order optimization is that even though it minimizes the training loss L_{train} , it dismisses the higher-order information such as curvature which correlates with the generalization, leading to a higher generalization error in test loss L_{test} according to [6]. Therefore, motivated by Chen *et al.* [6], the Sharpness-Aware Minimization (SAM) designed by Foret *et al.* [14], a second-order optimization technique, is executed to improve the generalization of our model for robust depression estimation in different scenarios.

Intuitively, SAM seeks to find the weight parameter w of a model whose entire neighbors in the range ρ have low training loss L_{train} compared with other weight parameters, as stated by Chen *et al.* [6]. This interpretation could be formulated into a minimax decision shown below:

$$\min_w \max_{\|\varepsilon\|_2 \leq \rho} L_{train}(w + \varepsilon), \quad (10)$$

which is a second-order problem. However, due to the complexity of solving the exact inner maximization with the optimum ε_{opt} , Foret *et al.* [14] employ the first-order approximation for better efficiency of calculating the sharpness aware gradient $\hat{\varepsilon}(w)$, which can be structured as:

$$\begin{aligned} \hat{\varepsilon}(w) &= \arg \max_{\|\varepsilon\|_2 \leq \rho} L_{train}(w) + \varepsilon^T \nabla_w L_{train}(w) \\ &= \rho \nabla_w L_{train}(w) / \| \nabla_w L_{train}(w) \|_2. \end{aligned} \quad (11)$$

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

$$w' = \nabla_w L_{train}(w) |_{w+\hat{\varepsilon}(w)}. \quad (12)$$

4 Experiments

In this study, we seek to model sequences of interactions to estimate the depression severity of each individual. Extensive experiments on DAIC-WOZ dataset have been conducted and the overall experimental methodology is to first train each single-modal model, including audio-, visual, and textual data, for the purpose of retrieving weights from effective feature extractors and then applying transfer learning to various multi-modal models.

4.1 Dataset

Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset [10, 17] contains clinical interviews of 189 participants designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder (PTSD). During each interview, several data in different format as well as modalities are recorded simultaneously. However, only

the acoustic recordings, facial key points, gaze directions, and transcriptions are chosen in this work, representing 3 different input data domains, namely audio (A), visual (V), and text (T). Moreover, the given GT is an eight-item Patient Health Questionnaire depression scale (PHQ-8), which indicates the severity of depression. A PHQ-8 Score ≥ 10 implies that the participant is undergoing a MD [26]. Although the DAIC-WOZ dataset [10] abounds in various data types and features, it contains assorted errors and problems, *e.g.*, small-scale dataset, imbalanced dataset, and labeling errors. These issues will potentially sabotage the model performance and mislead the model’s attention. Therefore, several techniques are applied to alleviate such burdens, such as sliding window technique, gender balancing (GB), weighted random sampler in PyTorch [33] etc.

4.2 Effectiveness of Different Fusion Methods

During the multi-modal training, we focus on two aspects: the impact pertains to different multimodalities and the effectiveness of the individual fusion approaches. For multimodality, we conduct experiments based on (1) AVT-modality and (2) AV-modality. As for the fusion approaches, in total, eight different fusion methods have been tested, which could be categorized into the traditional and weighting fusion method as listed below:

- Multiplication method: $\mathbf{y}_d^{multi} = \mathbf{x}_d^1 \otimes \mathbf{x}_d^2 \otimes \cdots \otimes \mathbf{x}_d^n$,
- Concatenation method: $\mathbf{y}_{n,d}^{cat} = f^{cat}(\mathbf{x}_d^1, \mathbf{x}_d^2, \dots, \mathbf{x}_d^n)$,
- Median method: $\mathbf{y}_d^{median} = median(\mathbf{x}_d^1, \mathbf{x}_d^2, \dots, \mathbf{x}_d^n)$,
- Maximum method: $\mathbf{y}_d^{max} = max(\mathbf{x}_d^1, \mathbf{x}_d^2, \dots, \mathbf{x}_d^n)$,
- Summation method: $\mathbf{y}_d^{sum} = \mathbf{x}_d^1 + \mathbf{x}_d^2 + \cdots + \mathbf{x}_d^n$,
- Mean method: $\mathbf{y}_d^{mean} = (\mathbf{x}_d^1 + \mathbf{x}_d^2 + \cdots + \mathbf{x}_d^n)/n$,
- Attentional fusion method,
- Sub-attentional fusion method,

where n and d are the number and the dimension of extracted feature vectors (\mathbf{x}). Furthermore, the attentional fusion method resembles the sub-attentional fusion method. The major difference is the number of attentional fusion layers. While the sub-attentional fusion method has individual attentional fusion layer for each of the 8 subclasses, only one single shared attentional layer has been utilized in the attentional fusion method.

The results of different fusion methods are summarized in Table 1. On both AV- and AVT-modality, the attention-based reweighting fusion methods generally perform better than the traditional ones with a 1% accuracy improvement on average. It indicates that an extra training layer for attentional feature fusion does provide advantages in harvesting deeper underlying depression cues for a better depression estimation. Moreover, forming the AVT-modality by adding textual data can consistently improve the accuracy of most fusion methods. Our sub-attentional fusion with AVT achieves the best score with 82.65% of accuracy while showing a satisfied f1-score with 0.65.

Table 1. Experimental results for multi-modal feature fusion with ConvBiLSTM as backbone. It demonstrates the effectiveness between different fusion methods and modalities.

Fusion method	Accuracy %		F1-Score	
	AV	AVT	AV	AVT
Multiplication	79.80	80.41	0.63	0.58
Concatenation	79.80	80.82	0.62	0.57
Median	80.20	82.04	0.63	0.59
Maximum	80.82	81.22	0.59	0.59
Summation	81.43	81.22	0.61	0.63
Mean	81.22	81.63	0.56	0.60
Attention	82.04	82.25	0.61	0.66
Sub-attention	82.04	82.65	0.58	0.65

Table 2. Experimental results of effectiveness before and after applying SAM, BiLSTM, and MUSDL, demonstrated with audio (A) and visual (V) modality.

SAM	BiLSTM	MUSDL	Accuracy %		F1-Score	
			A	V	A	V
✗	✗	✗	75.31	70.20	0.53	0.52
✓	✗	✗	76.12	76.73	0.56	0.54
✓	✓	✗	76.53	76.94	0.58	0.59
✓	✓	✓	76.73	79.59	0.61	0.61

4.3 Ablation Studies

To demonstrate how SAM, BiLSTM, and MUSDL reinforce the performance and have a better understanding of how our models estimate depression between both genders and participants, three following ablation studies have been carried out.

Effectiveness of Applying SAM, BiLSTM, and MUSDL. Ablation experiments are conducted regarding the using of SAM, BiLSTM and MUSDL in Table 2. An incremental performance gain regarding either F1-score or Accuracy is shown by the results demonstrating the efficacy of each individual component of our model, while the combination utilization of all these three techniques shows the best performance regarding audio and visual modality with 76.73% and 79.59% for accuracy, and 0.61 and 0.61 for F1-score respectively.

Sensitivity of Gender Depression Estimation. The purpose of the gender analysis is to dive deep into each modal and comprehend how sensitive each model is in terms of detecting MD between each gender and how significant the Gender Balancing (GB) technique is to suppress the gender bias phenomenon.

Table 3. Experimental results for the analysis of gender bias for depression estimation considering different model structures.

Model name	Modality	Accuracy %				F1-Score			
		Overall	Female	Male	Difference	Overall	Female	Male	Difference
ConvBiLSTM	A (No GB)	70.00	64.44	77.67	13.23	0.55	0.54	0.59	0.05
ConvBiLSTM	A	76.73	79.23	73.30	5.93	0.61	0.69	0.46	0.23
ConvBiLSTM	V	79.59	79.93	79.13	0.80	0.61	0.63	0.58	0.05
Atten ConvBiLSTM	AV	82.04	80.63	83.98	3.35	0.59	0.60	0.57	0.03
Sub-atten ConvBiLSTM	AVT	82.65	82.39	82.04	0.35	0.65	0.68	0.55	0.13

Therefore, the predicted test results of all clips are categorized into female and male groups, and their results are derived accordingly. In Table 3, all the results for the gender analysis are summarized with the best score marked in bold.

By observing the first two models, which are ConvBiLSTM with and without GB, one can notice a huge reduction of gender accuracy difference of around 7.3%. This signifies the seriousness of the role that the gender bias phenomenon plays in the acoustic features and how critical it is to handle it during the audio preprocessing stage. Visual features, on the other hand, show no problem of gender bias with a gender accuracy difference of less than 1%, which is also understandable as one can imagine how challenging it is for a person to distinguish a participant's gender solely based on the 68 3D facial key points.

Furthermore, a decreased tendency of gender accuracy difference in multi-modal model can be discovered, implying that the more different modalities are fused, the lower this acoustic gender bias phenomenon shows up. This is the fact that by fusing variant data modalities, the model can learn diverse feature from different input sources and balance the gender bias. Finally, with the Sub-attentional ConvBiLSTM model trained on AVT modality, the lowest gender accuracy difference 0.35% is achieved, and thus it has the highest sensitivity in gender depression estimation over 82% accuracy in both genders. The proposed Sub-attentional ConvBiLSTM model also shows the best performance regarding F1-score overall as 0.65, however, the gender difference regarding F1 score is still a limitation of our model and thereby a future research direction.

Sensitivity of Participants Depression Estimation. To further allay the concern regarding the representation of our models since the depression of a participant in GT is diagnosed by the specialist based on a whole interview instead of a clip, the participant analysis is conducted by recombining the clips as well as the predicted scores back into each participant to form the original interview. The final PHQ-8 Score for each participant is then computed as the mean of all clips, and a threshold of 0.5 is set for the final PHQ-8 Binary, meaning that if over 50% of the clips of the current participant is being classified as depressed by the multi-modal model, it can be concluded that this participant is having MD, and vice versa. An illustration is demonstrated in Fig. 4 and the final results are summarized in Table 4.

For the normal case, the Sub-attentional model predicts the whole clips from the interview of the participant as class 1 or 0, which is shown in the first two

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

Fig. 4. A visualization of the analysis on participant-level. Under each clip, the PHQ-8 Subscore and PHQ-8 Binary are denoted with the value and color inside the block, respectively. Gray background: class 1, White background: class 0.

Table 4. Experimental results of different model structures for participant-level depression estimation.

Model name	Modality	Accuracy %			F1-Score		
		Clipped data	Participant-based	Improvement	Clipped data	Participant-based	Improvement
ConvBiLSTM	A (No GB)	70.00	70.21	0.21	0.55	0.53	-0.02
ConvBiLSTM	A	76.73	78.72	1.99	0.61	0.64	0.03
ConvBiLSTM	V	79.59	78.72	-0.87	0.61	0.62	0.01
Atten ConvBiLSTM	AV	82.04	80.85	-1.19	0.59	0.61	0.02
Sub-atten ConvBiLSTM	AVT	82.65	85.11	2.46	0.65	0.70	0.05

participants in Fig. 4. For the rest of the specific situations, one can notice a mix of predicted classes for the clips in each interview. This is due to the fact of inconsistent expression of depressive symptoms throughout the whole interview despite having MD, which causes non-error mistakes, and some ambiguous clips, which confuse the model. This mix, however, can be rectified through the analysis as one can observe from the final binary status of MD in Fig. 4. Overall, the tolerance between both clipped data and participant-based accuracies is relatively low, less than 3% according to Table 4. Therefore, it concludes that all of the models, as well as the technique of training on the clipped dataset for depression estimation, are valid and representative. Furthermore, one can perceive that there is even a performance improvement of around 2.5% accuracy and 0.05 F1 score in our best model, Sub-attentional ConvBiLSTM.

4.4 Automatic Depression Estimation

To compare with the state-of-the-art approaches, the following scores are further derived: F1-Score, Precision, Recall, MAE, and RMSE, shown in Table 5. Here, the single- and multi-modal models are both included, along with different analysis approaches, namely clipped data-based as well as participant-based marked with †. Moreover, the model reproduction results (*) of the baselines [2, 29], which are trained on our generated dataset, are also included. The best scores in our methods and previous works are both marked in bold. One major

Table 5. A comparison with the state-of-the-art methods. We assess two outcomes: (1) binary status of having MD and (2) the level of depression severity.

Comparison of SOTA		PHQ-8 binary			PHQ-8 score		
Method	Modality	PL	F1-Score	Precision	Recall	MAE	RMSE
Previously published works							
Ma <i>et al.</i> [29]	A	L	0.52	0.35	1.00	-	-
Valstar <i>et al.</i> [44]	A	H	0.46	0.32	0.86	5.36	6.74
Williamson <i>et al.</i> [47]	V	H	0.53	-	-	5.33	6.45
Valstar <i>et al.</i> [44]	V	H	0.50	0.60	0.43	5.88	7.13
Alhanai <i>et al.</i> [2]	AT	M	0.77	0.71	0.83	5.10	6.37
Valstar <i>et al.</i> [44]	AV	H	0.50	0.60	0.43	5.52	6.62
Gong <i>et al.</i> [16]	AVT	H	0.70	-	-	2.77	3.54
Comparable baselines (selected previous works with a data processing pipeline comparable to ours)							
* Ma <i>et al.</i> [29]	A	L	0.48	0.38	0.65	-	-
* Alhanai <i>et al.</i> [2]	AT	M	0.44	0.29	0.93	5.92	7.68
Our approaches							
ConvBiLSTM	A	L	0.61	0.56	0.66	5.19	6.93
ConvBiLSTM	V	L	0.61	0.64	0.58	6.17	8.06
Atten ConvBiLSTM	AV	L	0.59	0.79	0.47	4.92	5.86
† Atten ConvBiLSTM	AV	L	0.61	0.78	0.50	5.06	6.06
Sub-atten ConvBiLSTM	AVT	L	0.65	0.73	0.58	4.99	6.67
† Sub-atten ConvBiLSTM	AVT	L	0.70	0.89	0.57	5.04	6.98

† Participant-based analysis PL: Preprocessing Level H: High M: Medium L: Low

* Model reproduced on the DAIC-WOZ dataset with our preprocessing pipeline

difference between our approach and prior works is that our approach does not heavily rely on complex feature engineering techniques. We utilize raw input modalities from audio, visual, and text data to realize automatic depression estimation as it is more public-friendly and can potentially improve diagnostic availability, whereas previous works use engineered features such as topic modeling context [16], Question/Answer pair [47], vocal tract resonances [47], and MFCCs [2, 44, 47]. However, our models still achieve highly comparable results.

5 Conclusion

In this work, we proposed a novel multi-modal deep-learning-based approach, *i.e.*, Sub-attentional ConvBiLSTM, to achieve end-to-end depression estimation while using less prepossessing techniques. By leveraging multi-modal data with such a hierarchical model structure to capture the short- and long-term temporal as well as spectral features, Sub-attentional ConvBiLSTM has demonstrated great success in harvesting deeper underlying depression cues and thus achieves an exceptional performance with 85.11% for accuracy, 0.89 for precision, 0.70 for f1-score, which outperforms our baseline, *i.e.*, DepAudioNet [29], by a large margin. Furthermore, the proposed gender balancing technique has also been proven to have a strong effect on alleviating gender bias issue in acoustic features.

Finally, the participant-level analysis justifies the efficacy of our model trained on the clipped dataset and leveraging sliding windows during participant-level test. In conclusion, our method has competitive performance with current existed approaches for depression estimation using the knowledge from audio, visual, and text modalities while considering imbalanced, gender bias and small-scale dataset problems, which ensures the efficiency of depression estimation.

Acknowledgement. The research leading to these results was supported by the SmartAge project sponsored by the Carl Zeiss Stiftung (P2019-01-003; 2021–2026).

References

1. Akbar, H., Dewi, S., Rozali, Y.A., Lunanta, L.P., Anwar, N., Anwar, D.: Exploiting facial action unit in video for recognizing depression using metaheuristic and neural networks. In: ICCSAI (2021)
2. Al Hanai, T., Ghassemi, M.M., Glass, J.R.: Detecting depression with audio/text sequence modeling of interviews. In: Interspeech (2018)
3. Al Jazaery, M., Guo, G.: Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Trans. Affect. Comput.* **12**(1), 262–268 (2021)
4. Bhukya, B.B., Sravanthi, K.: Major depression disorder (2019)
5. Chen, Q., Chaturvedi, I., Ji, S., Cambria, E.: Sequential fusion of facial appearance and dynamics for depression recognition. *Pattern Recognit. Lett.* **150**, 115–121 (2021)
6. Chen, X., Hsieh, C.J., Gong, B.: When vision transformers outperform ResNets without pre-training or strong data augmentations. In: ICLR (2022)
7. Cohn, J.F., et al.: Detecting depression from facial actions and vocal prosody. In: ACII (2009)
8. Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., Epps, J.: Diagnosis of depression by behavioural signals: a multimodal approach. In: AVEC@ACM Multimedia (2013)
9. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K.: Attentional feature fusion. In: WACV (2021)
10. DAIC-WOZ Database. <https://dcapswoz.ict.usc.edu/>. Accessed 21 Oct 2019
11. Dinkel, H., Wu, M., Yu, K.: Text-based depression detection on sparse data. arXiv preprint [arXiv:1904.05154](https://arxiv.org/abs/1904.05154) (2019)
12. Du, Z., Li, W., Huang, D., Wang, Y.: Encoding visual behaviors with attentive temporal convolution for depression prediction. In: FG (2019)
13. Dumpala, S.H., Rempel, S., Dikaios, K., Sajadian, M., Uher, R., Oore, S.: Estimating severity of depression from acoustic features and embeddings of natural speech. In: ICASSP (2021)
14. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. In: ICLR (2021)
15. Fossi, L., Faravelli, C., Paoli, M.: The ethological approach to the assessment of depressive disorders. *J. Nerv. Mental Dis.* **172**(6), 332–341 (1984)
16. Gong, Y., Poellabauer, C.: Topic modeling based multi-modal depression detection. In: AVEC@ACM Multimedia (2017)
17. Gratch, J., et al.: The distress analysis interview corpus of human and computer interviews. In: LREC (2014)

18. Guo, Y., Zhu, C., Hao, S., Hong, R.: Automatic depression detection via learning and fusing features from visual cues. arXiv preprint [arXiv:2203.00304](https://arxiv.org/abs/2203.00304) (2022)
19. Halfin, A.: Depression: the benefits of early and appropriate treatment. Am. J. Manag. Care **13**(4), S92 (2007)
20. Hao, Y., Cao, Y., Li, B., Rahman, M.: Depression recognition based on text and facial expression. In: SPIE (2021)
21. He, L., Jiang, D., Sahli, H.: Multimodal depression recognition with dynamic visual and audio cues. In: ACII (2015)
22. He, L., et al.: Deep learning for depression recognition with audiovisual cues: a review. Inf. Fusion **80**, 56–86 (2022)
23. Jacobi, F., et al.: Prevalence, co-morbidity and correlates of mental disorders in the general population: results from the German health interview and examination survey (GHS). Psychol. Med. **34**(4), 597–611 (2004)
24. Joshi, J., et al.: Multimodal assistive technologies for depression diagnosis and monitoring. J. Multimodal User Interfaces **7**(3), 217–228 (2013)
25. Kaya, H., Salah, A.A.: Eyes whisper depression: a CCA based multimodal approach. In: ACM Multimedia (2014)
26. Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B., Berry, J.T., Mokdad, A.H.: The PHQ-8 as a measure of current depression in the general population. J. Affect. Disord. **114**(1–3), 163–173 (2009)
27. Kupfer, D.J., Frank, E., Perel, J.M.: The advantage of early treatment intervention in recurrent depression. Arch. Gen. Psychiatry **46**(9), 771–775 (1989)
28. Lin, C.J., Lin, C.H., Jeng, S.Y.: Using feature fusion and parameter optimization of dual-input convolutional neural network for face gender recognition. Appl. Sci. **10**(9), 3166 (2020)
29. Ma, X., Yang, H., Chen, Q., Huang, D., Wang, Y.: DepAudioNet: an efficient deep model for audio based depression classification. In: AVEC@ACM Multimedia (2016)
30. World Health Organization: Depression and other common mental disorders: global health estimates. Technical report, World Health Organization (2017)
31. Pampouchidou, A., et al.: Depression assessment by fusing high and low level features from audio, video, and text. In: AVEC@ACM Multimedia (2016)
32. Pampouchidou, A., et al.: Automatic assessment of depression based on visual cues: a systematic review. IEEE Trans. Affect. Comput. **10**(4), 445–470 (2019)
33. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: NeurIPS (2019)
34. Rathi, S., Kaur, B., Agrawal, R.: Enhanced depression detection from facial cues using univariate feature selection techniques. In: PReMI (2019)
35. Ray, A., Kumar, S., Reddy, R., Mukherjee, P., Garg, R.: Multi-level attention network using text, audio and video for depression prediction. In: AVEC@MM (2019)
36. Saidi, A., Othman, S.B., Saoud, S.B.: Hybrid CNN-SVM classifier for efficient depression detection system. In: IC_ASET (2020)
37. Salimath, A.K., Thomas, R.K., Reddy, S.R., Qiao, Y.: Detecting levels of depression in text based on metrics. arXiv preprint [arXiv:1807.03397](https://arxiv.org/abs/1807.03397) (2018)
38. Sardari, S., Nakisa, B., Rastgoo, M.N., Eklund, P.: Audio based depression detection using convolutional autoencoder. Expert Syst. Appl. **189**, 116076 (2022)
39. Scherer, K.R.: Vocal affect expression: a review and a model for future research (1986)
40. Stepanov, E.A., et al.: Depression severity estimation from multiple modalities. In: HealthCom (2018)

41. Tang, Y., et al.: Uncertainty-aware score distribution learning for action quality assessment. In: CVPR (2020)
42. Uddin, M.A., Joolee, J.B., Sohn, K.A.: Deep multi-modal network based automated depression severity estimation. IEEE Trans. Affect. Comput. (2022)
43. Valstar, M., et al.: AVEC 2016: depression, mood, and emotion recognition workshop and challenge. In: ACM Multimedia (2016)
44. Valstar, M., et al.: AVEC 2016: depression, mood, and emotion recognition workshop and challenge. In: AVEC@ACM Multimedia (2016)
45. Waxer, P.: Nonverbal cues for depression. *J. Abnorm. Psychol.* **83**(3), 319 (1974)
46. WHO: Depression key facts. World Health Organization (2021). <https://www.who.int/news-room/fact-sheets/detail/depression>
47. Williamson, J.R., et al.: Detecting depression using vocal, facial and semantic communication cues. In: AVEC@ACM Multimedia (2016)
48. Williamson, J.R., Quatieri, T.F., Helfer, B.S., Horwitz, R., Yu, B., Mehta, D.D.: Vocal biomarkers of depression based on motor incoordination. In: AVEC@ACM Multimedia (2013)
49. Xie, W., et al.: Interpreting depression from question-wise long-term video recording of SDS evaluation. *IEEE J. Biomed. Health Inform.* **26**(2), 865–875 (2022)
50. Zhu, Y., Shang, Y., Shao, Z., Guo, G.: Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Trans. Affect. Comput.* **9**(4), 578–584 (2018)