Depression Severity Level Classification Using Multitask Learning of Gender Recognition

Yang Liu , Xiaoyong Lu * ,Daimin Shi , and Jingyi Yuan
Northwest Normal University Lanzhou, Gansu, China
E-mail: luxy@nwnu.edu.cn

***Abstract*—Speech based classification of depression has been widely used. However, most classification studies focus on binary classification to distinguish depressed subjects from non-depressed subjects. In this paper, we describe the depression classification task as a severity classification problem to provide finer grained classification results. we formulate the Attention deep learning network for Speech Depression Recognition (SDR) using the Mel-frequency cepstral coefficient (MFCC) features as the input. The attention along with the convolutional neural network and the bidirectional long short-term memory network (CNN-BLSTM) embedding jointly attends to information from different representations of the same MFCC input sequence. The CNN-LSTM embedding helps in attending to the dominant depression features by identifying positions of the features in the sequence. In addition to Attention and CNN-LSTM embedding, we apply multi-task learning with gender recognition as an auxiliary task. The auxiliary task helps in learning the gender-specific features that influence the depression characteristics in speech and results in improved accuracy of Speech Depression Recognition, the primary task. We conducted all our experiments on Depression dataset. We can achieve an overall F1sorce of 81.5% and average class accuracy of 89.3%, on SDR for depression classes.**

I. INTRODUCTION

Depression is a mental illness that robs people of their self-esteem and enjoyment in life. More significantly, it has the potential to lead to irrational behavior. The Top Ten Most Serious Diseases in 2002 [1] According to relevant research, college students are affected by variables such as studies and graduation, and their mental health problems, such as depression and anxiety, become more evident [2-5]. Based on the results of the survey, the researchers estimated that 16.6 percent of Chinese adults had experienced mental illness at some point in their lives, a much higher rate than in previous surveys, which were limited in scope. Anxiety disorders were the most common. Also on the rise was depression, which had affected 6.9 percent of those surveyed throughout their lives and 3.6 percent in the previous 12 months [6- 8]. but only a small number of them have been found and treated [9, 10]. Many patients are unable to receive prompt therapy in the early (Depression tendency), leading to the aggravation of the disease (Depression). The most common detection methods for depression rely on the use of scales for patients and the clinical experience of doctors [11]. In addition, doctors spend a lot of energy in the process, which can increase the probability of misdiagnosis in subsequent work. Thus, it is necessary to develop a method to assist doctors.

Psychological studies have observed differences in language use between depressed and non-depressed patients [12-14]. Based on these facts, researchers studied the use of different machine learning in combination with speech alone [15,16], and combination [17,18]. These techniques use, for example, support vector machines, convolution neural networks, and short-term and long-term memory networks. However, these methods only consider speech characteristics, but they have less influence on categories that are difficult to distinguish, such as gender, age, and occupation.

The gender of the speaker has a significant effect on automatic speech depressive recognition and its impact has been studied over the years [19-21]. [21] shows that some of the input features for SDR such as vowel-level formant, F0 features have different values for each gender. In [20], the authors propose a method where the gender effect assessment is evaluated by comparing the Depressed/Non-Depressed accuracies for both genders. The results showed that there were differences among different groups. Researcher M. Muzammel noticed that the final accuracy rate of the deep learning model was different in the experiment of different gender individuals, which indicated that the deep learning model could not further distinguish the influence of some characteristics on gender in the training process.

Introducing gender factors into depressive predisposition testing as a breakthrough in the difficulty of categorizing features has proved to be a good way to alleviate these two problems. In this article, we present a multitask model based on gender as a secondary task. First, the MFCC of the speech segment is obtained, and then the extracted features are trained end-to-end using a multitask model that uses CNN-LSTM to explore the temporal and. spatial characteristics. Thirdly, we propose a multi-task learning approach with gender recognition as an auxiliary task to improve the effectiveness of the task in some categories. Finally, we classify depressive tendencies using a fully connected layer. Compared with previous methods, our gender-based task-assisted multitask model is superior to the one without multitask.

The rest of this paper is structured as follows: Section II, introduces the method of Deep Learning Model structure, and Section III details the experiments performed and analyses the results obtained, while Section IV. summarizes the article and discusses future research on the subject.

* Corresponding author.

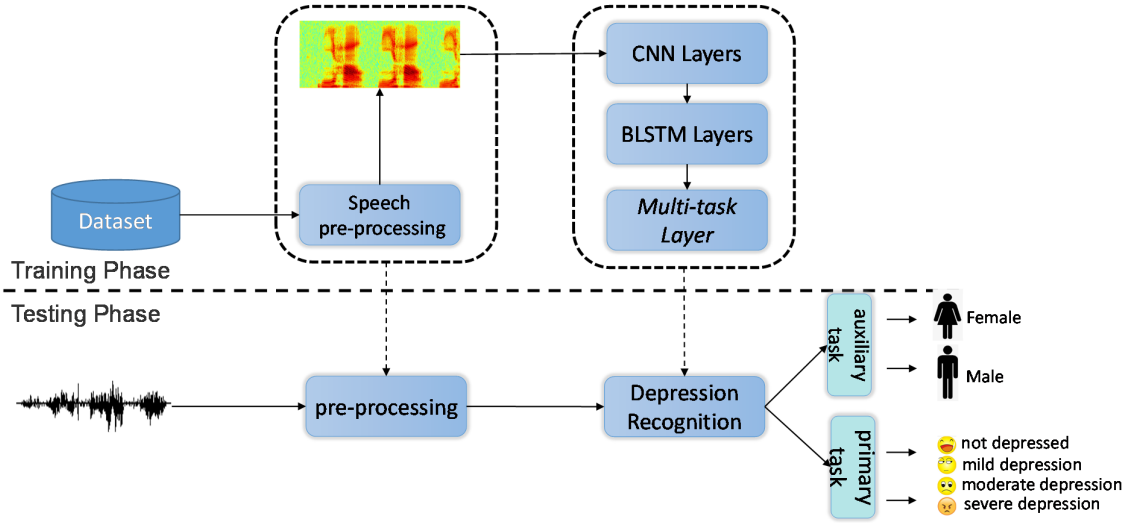

Fig 1. The framework of the proposed speech depression recognition system

II. METHOD

A. Overview

Fig 1 shows the overall experimental flow. For the training phase, the main components are speech pre-processing, depressive tendency model construction, and multi-task classification. Firstly, pre-processing of the speech signal is used, followed by feature extraction of the speech, then feeding into the designed depressive tendency recognition model which consists of CNN network layer, BLSTM network layer, attention mechanism layer, fully connected layer, and finally the features are fed into the model for training in a multi-tasking manner and the optimal training parameters are saved. In the testing phase, the speech fragments are also pre-processed, features are extracted and the processed features are then sent to the model for training in a multi-task manner. The speech fragments are preprocessed and feature extracted in the same way as the training phase, and the processed features are fed into the trained and best-preserved model for depressed tendencies prediction.

B. Feature Extraction

The audio files were sampled at 16 kHz and stored in a 16-bit signed wav format. The voice signal was pre-processed with CoolEdit and Praat to remove noisy parts, such as coughing or throat clearing, and parts where the operator interfered with the subject, such as correcting errors or providing task-related details. We set 6 seconds as the audio input length. Any audio file longer than 6 seconds was truncated to 6 seconds. Files shorter than 6 seconds were padded with zeros. In all cases, features were obtained using a short-time Fourier transform of length 512, using a Hamming window of 20ms with 50% overlap. MFCCs were calculated for each frame and the number of filter banks was set to 64. MFCCs features were generated using wav files and a python speech feature library. The 380 matrices generated were fed into the model, where 380 was considered to be the sequence length. The proposed method is shown in Fig. 2.

Fig 2. The extraction process of MFCCs

C. CNN-BLSTM Embedding

The information of voice signals and text can be approximated as sequence information, but the semantic content of voice signals is much greater than that of text. So instead of using simply BLSTM, we use MFCC to express the vector of voice signal that is fed into CNN-BLSTM network (CBL). As the feature sequence is fed into the CNN layer, each sub-module includes convolution, batch normalization, average pool, and discard operations to produce the vector. BLSTM then generates an advanced feature sequence by learning the context dependence between long-term temporal features before and after each CNN-operated vector. For each of the BLSTM vectors, the BLSTM is forward and backward linked. The proposed method is shown in Fig. 3.

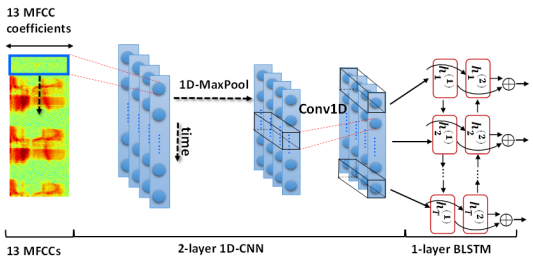

Fig 3. The extraction process of CNN-BLSTM Embedding

D. Attention

With limited computing power, Attention Mechanism, as a resource allocation scheme, uses limited computing resources to process more important information and is the main means to solve the problem of information overload. When using the neural network to process a large amount of input information, it can also refer to the human brain's attention mechanism and select only some key information inputs for processing to improve the efficiency of the neural network.

Following the BLSTM layer, an attention network aggregates information from the BLSTM hidden states $H^{blstm}$and get a fixed-length vector Z as the encoding of the speech segment. For each vector h_i^{blstm} in a sequence of inputs $H^{blstm}=\{h_1^{blstm},h_2^{blstm},\ldots,h_{L/w}^{blstm}\}$, the attention weights α_t are given by:

$$\alpha_t=\frac{\exp(f(h_t^{blstm}))}{\sum_{J=1}^{T}\exp(f(h_t^{blstm}))} \quad (1)$$

where $f(h_t^{blstm})$ is defined using the trainable parameters W as follows:

$$f(h_t^{blstm})=\tanh(W^T h_t^{blstm}) \quad (2)$$

Then we summary all the weighted sums to get the final encoding vector $z \in \mathbb{R}^{2d_{lstm}}$

$$h_j^{atten}=\alpha_t H^{blstm} \quad (3)$$

$$z=h_1^{atten}+h_2^{atten}\cdots h_j^{atten} \quad (4)$$

E. Multi-task

Multi-task learning (MTL) is a branch of machine learning. Its training process includes multiple learning tasks. By exploiting the similarities and differences between different tasks, the generalization ability and prediction accuracy of the model can be improved. To achieve multi-task learning, parameter sharing between models needs to be implemented. Thus, a multi-task learning model is a combination of multiple overlapping machine learning models. We thus propose an MTL-Attention network where gender recognition is the only auxiliary task to improve the main SER task. In order to better the fused features, we designed the Attention, specifically as shown is that for the fused features each neuron contains all the feature information, but not every fusion is valid, so the significance of using Attention for multi-task learning by adding attention to select useful neurons is that each layer of fusion works on a specific set of features of the input, thus allowing the model to learn multiple representations common to both tasks in a more fine-grained space. In addition, inspired by [22-25], we pay special attention to Muti-Layer Perception (MLP) and explore MLP in detail. For the MLP layer, we correct the attention layer of the merged network, which is similar to the attention method in the previous section. The formula is as follows Eq. (1) (2) (3). The proposed method is shown in Fig. 4.

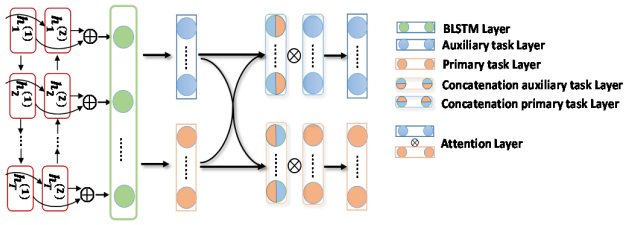

Fig 4. The extraction process of Multi-task

F. Multi-Task Training

we adopt a joint model to consider the two tasks and update parameters by joint optimizing. The model is optimized by the following objective function.

$$\mathcal{L}=\alpha\mathcal{L}_{depression}+\beta\mathcal{L}_{gender} \quad (5)$$

Where $\mathcal{L}_{depression}$and $\mathcal{L}_{gender}$ are the losses for depression classification and gender classification, respectively. α and β represent the weights for the two tasks. Use cross-entropy as the loss function for both tasks. We tried several weights for the network to the main task and found that setting both weights to 1.0 produced the best accuracy of depression classification.

III. EXPERIMENT

A. Speech data

The depression dataset [26] was collected from Chinese students at the university. All participants are Chinese adults with the age range from 18 to 25. Each participant was told to fill in basic personal information after signing the informed consent form and begin the experiment under the guidance of the operator. The recordings were recorded at 44,1 kHz with 16-bit sample rate. During the collection of speech recordings, speakers covered different degrees of severity of depression from healthy to severely depressed. Two clinicians (CL)-rated depression assessment scales. The Hamilton Rating Scale for Depression [27] (HAMD) and the Beck Depression Index [28] (BDI) were provided to define the severity of depression (Table I). which is currently one of the most common methods for early warning of depression in clinical practice. In the 4-level severity classification task, data from levels 5 and 3-4 were combined into 'severe' and 'moderate' levels, respectively. Data from level 1 were used for the 'normal' level.

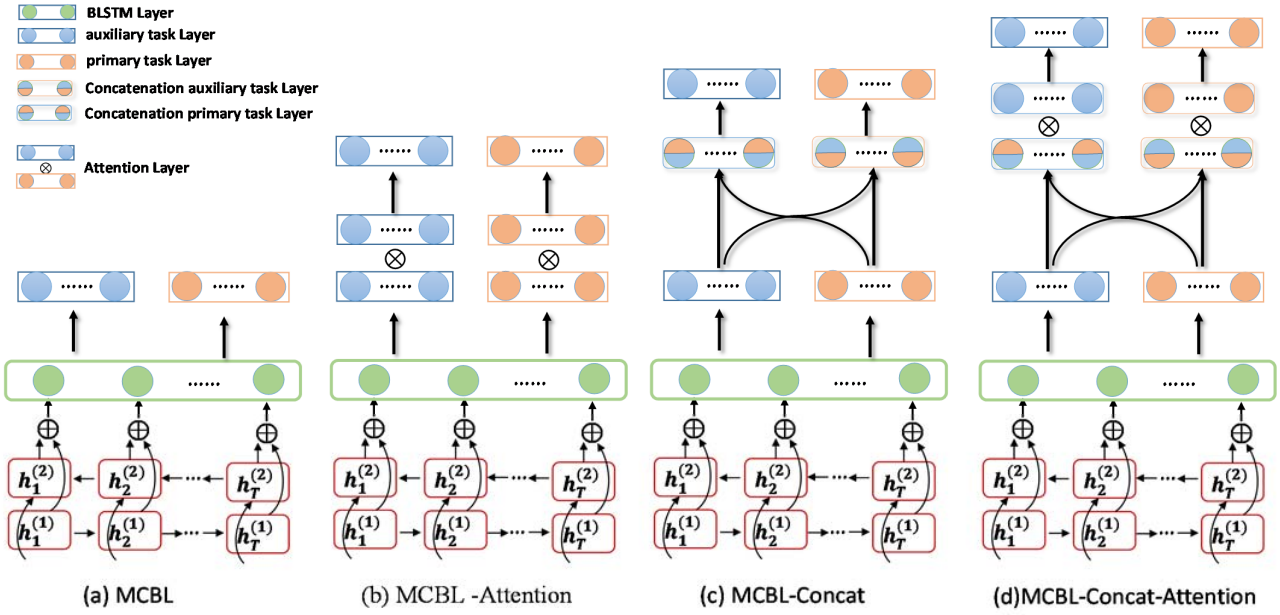

Fig 5. Overview of the model architecture.

TABLE I. DEPRESSION SEVERITY CATEGORIES

Category	HAM-D score	BDI score
1.not depressed	0-7	0-13
2.mild depression	8-13	14-19
3.moderate depression	14-18	20-28
4. severe depression	19-22	29-63
5.very severe depression	23-52	

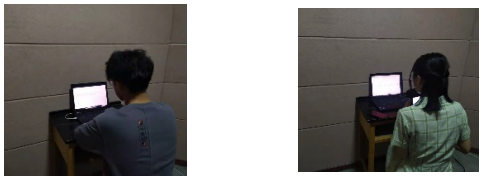

Fig 6. The collect speech process

B. *Experimental Setting*

We implemented the depressive tendency classification models using Pytorch deep learning library version 1.2.0. The models were trained using the NVIDIA Tesla K40C graphical processing unit (GPU). We used the Adam optimization algorithm [29] to optimize the parameters in our model and adopted the suggested hyper-parameters for optimization. We used 100 epochs for training and saved the best one as the final model. The initial learning rate was set to 0.001 and the batch size was set to 16. We applied a dropout after the BLSTM layer with 0.2 dropout probability. The dimensions mentioned in Section 3 are summarized in Table II

TABLE II. DIMENSION DETAILS.

Attribute	Value
Number of MFCC features	13
Input Sequence Length	380
Number of convolution filters	256
Number of LSTM hidden units	64
Number of Attention Layers	128
Number of auxiliary task Layers	64
Number of primary task Layers	64
Number of concatenation Layers	128

C. *Depression Recognition on Dataset*

We used the mean and F1 scores as evaluation criteria for our experiments. In Table III, the results show that the multitasking attention mechanism is feasible for the recognition of depression and that the use of MTL still improved the model's recognition of depression. As can be seen in Table III, the addition of MTL increases the accuracy to89% and the F1score to 81%, which is better than without the use of MTL. Thus, the multitask learning multitask attention model with gender recognition as a secondary task performed best.

TABLE III. RESULT OF MODELS

Methods	Class Accuracy	Class F1 Score
CNN [30]	0.759	0.679
LSTM [30]	0.667	0.535
CBL (our)	0.761	0.674
MCBL-Concat-Attention (our)	0.893	0.815

D. *Ablation Study*

We experimented based on the basic model MTL-CNN-LSTM (MCBL) and kept the other components unchanged. First, we use a multitask approach called MCBL -Attention. The results show that the MCBL -Attention approach in Table IV is 72.4%. We observed a decline in the performance of all indicators in the depression dataset. Next, we use a feature fusion method called MCBL-Concat. The results show that the method based on feature fusion in Table IV is 72%. We observed an indicator in the performance of all indicators in the depression dataset. Finally, we add attention to feature fusion, which is called MCBL-Concat-Attention. The results show that the MCBL-Concat-Attention method in Table IV is 89.3%. We observed an increase in the performance of all indicators in the depression dataset. We think our method can classify the fused feature information very well. The MCBL method is shown in Fig. 5.

TABLE IV. ABLATION STUDY ON DATASETS

Methods	Class Accuracy	Class F1 Score
MCBL	0.779	0.662
MCBL -Attention	0.724	0.559
MCBL-Concat	0.720	0.540
MCBL-Concat-Attention	0.893	0.815

E. *Visualization*

In the attempt to better understand what the MCBL-Concat-Attention-MTL model has learned, we visualize the feature distribution by t-SNE [31] to further analyze our MCBL-Concat-Attention-MTL method. Fig. 4 depicts the embedding deep features of the testing set on 2-dimensional space. Fig. 7(a) represents the scattered features from the CBL model, while all classes can cluster well and different classes can be pushed apart with our MCBL-Concat-Attention-MTL model in Fig. 4(b). These figures demonstrate that our MCBL-Concat-Attention-MTL method can classify accurately, especially the moderate depression class(yellow).

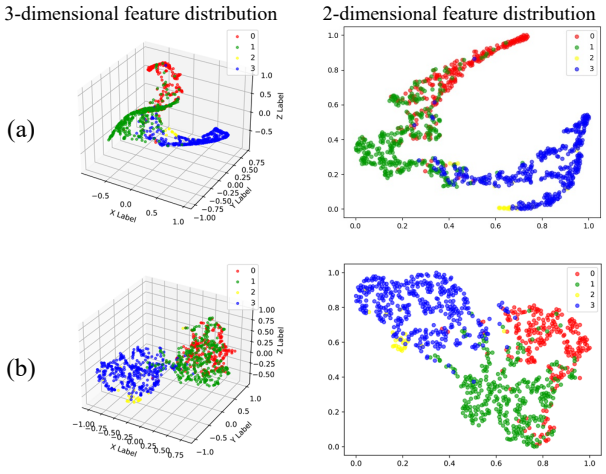

Fig 7. 2,3-dimensional feature distribution in the testing set. (a) is plotted under the CBL model, while (b) is plotted under our MCBL-Concat-Atten-MTL model. The categories are respectively marked by red (not depression), green (mild depression), yellow (moderate depression), and blue (severe depression).

F. *Average class accuracy*

As can be seen from the visual analysis above, multitasking has an effect on feature clustering, indicating that specific aspects are influenced by the gender factor, and we used the average accuracy to better examine each case. In Fig 6, we show the confusion matrix, which summarizes the classification's performance.

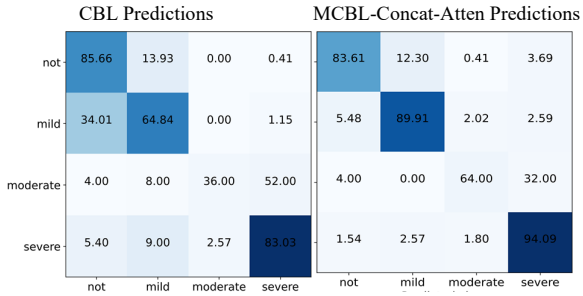

Fig 8. Confusion Matrix to show the average class accuracies

Initially, without multi-task learning, the validation accuracy reached 76% and the validation F1 score 67% (Table III). In particular, the difference between the CBL models was not very good. The attention model with multi-task learning resulted in a 28% increase in incorrectness for the moderate lesson. This can be attributed to the model's improved ability to identify gender-dependent features (or their relative importance). Thus, across the categories, there was a significant improvement in accuracy compared to the state-of-the-art results. The slight decrease in accuracy for the main depression score can be attributed to better generalization.

IV. CONCLUSION

In this paper, we borrow ideas from the attention mechanism to design an MCBL-Concat-Attention identification model for depressive tendencies classification. We conducted experiments on the database of college students' depression tendency, and the results show that the method using the attention mechanism is feasible.

This work can be extended to a system for developing a mobile phone program that can take advantage of the voice information available for automatic speech recognition. And we will analyze the complexity of the algorithm to improve the operation efficiency of embedded devices. Language features can reveal important information about language content related to the mental health of depressive patients.

ACKNOWLEDGEMENT

This research has received funding from the academic requirements for the National Science Foundation of China (NSFC) under grants No. 31860285 and No. 31660281. Additionally, part of this work is performed in the Scientific Research Project in Higher Education Institutions of Gansu Province (Grant No. 2017A-165). We also want to thank the reviewers for their thoughtful comments and efforts towards improvement.

REFERENCES

[1] W. H. Organization, Neurological disorders: public health challenges. World Health Organization, 2006.

[2] R. Beiter et al., "The prevalence and correlates of depression, anxiety, and stress in a sample of college students," vol. 173, pp. 90-96, 2015.

[3] A. K. Ibrahim, S. J. Kelly, C. E. Adams, and C. J. J. o. p. r. Glazebrook, "A systematic review of studies of depression prevalence in university students," vol. 47, no. 3, pp. 391-400, 2013.

[4] C. J. Othieno, R. O. Okoth, K. Peltzer, S. Pengpid, and L. O. J. J. o. a. d. Malla, "Depression among university students in Kenya: Prevalence and sociodemographic correlates," vol. 165, pp. 120-125, 2014.

[5] K. Peltzer, S. Pengpid, S. Olowu, and M. J. J. o. p. i. A. Olasupo, "Depression and associated factors among university students in Western Nigeria," vol. 23, no. 3, pp. 459-465, 2013.

[6] Y. Huang *et al.*, "Prevalence of mental disorders in China: a cross-sectional epidemiological study," vol. 6, no. 3, pp. 211-224, 2019. J. Wei and Z. J. C. J. H. P. Sang, "Research progress on family factors of depression for college students," vol. 25, pp. 1752-1756, 2017.

[7] X. J. E. R. Luo, "The Status Quo and Countermeasures of College Students' Mental Health Education," vol. 1, pp. 112-118, 2018.

[8] K. D. Michael, T. J. Huelsman, C. Gerard, T. M. Gilligan, M. R. J. C. Gustafson, and C. P. Journal, "Depression Among College Students: Trends in Prevalence and Treatment Seeking," vol. 3, no. 2, 2006.

[9] J. Kisch, E. V. Leino, M. M. J. S. Silverman, and L.-T. Behavior, "Aspects of suicidal behavior, depression, and treatment in college students: Results from the Spring 2000 National College Health Assessment Survey," vol. 35, no. 1, pp. 3-13, 2005.

[10] E. J. Miller and H. J. P. s. Chung, "A literature review of studies of depression and treatment outcomes among US college students since 1990," vol. 60, no. 9, pp. 1257-1260, 2009.

[11] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. J. S. C. Quatieri, "A review of depression and suicide risk assessment using speech analysis," vol. 71, pp. 10-49, 2015.

[12] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. J. I. t. o. B. E. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," vol. 47, no. 7, pp. 829-837, 2000.

[13] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in Proceedings of the 6th international workshop on audio/visual emotion challenge, 2016, pp. 35-42.

[14] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller, "Enhancing speech-based depression detection through gender dependent vowel-level formant features," in Conference on Artificial Intelligence in Medicine in Europe, 2017, pp. 209-214: Springer.

[15] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level Attention network using text, audio and video for Depression Prediction," in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 81-88.

[16] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017, pp. 53-59.

[17] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4438-4446.

[18] A. Vaswani et al., "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998-6008.

[19] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. J. a. p. a. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," 2019.

[20] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, and A. J. M. L. w. A. Othmani, "AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis," vol. 2, p. 100005, 2020.

[21] B. Vlasenko, H. Sagha, N. Cummins, and B. Schuller, "Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition," 2017.

[22] I. O. Tolstikhin *et al.*, "MLP-Mixer: An all-MLP Architecture for Vision," 2021.

[23] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. J. a. e.-p. Hu, "Beyond Self-attention: External Attention using Two Linear Layers for Visual Tasks," p. arXiv: 2105.02358, 2021.

[24] X. Ding, X. Zhang, J. Han, and G. Ding, "RepMLP: Re-parameterizing Convolutions into Fully-connected Layers for Image Recognition."

[25] L. J. a. p. a. Melas-Kyriazi, "Do you even need attention? a stack of feed-forward layers does surprisingly well on imagenet," 2021.

[26] X. Lu *et al.*, "Development of a Chinese Depressed Speech Corpus Based on The Disturbed Effect of Self-Processing," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 718-722: IEEE.

[27] M. Hamilton, "The Hamilton rating scale for depression," in *Assessment of depression*: Springer, 1986, pp. 143-152.

[28] C. L. Carter and C. M. J. J. o. A. Dacey, "Validity of the Beck Depression Inventory, MMPI, and Rorschach in assessing adolescent depression," vol. 19, no. 3, pp. 223-231, 1996.

[29] D. P. Kingma and J. J. a. p. a. Ba, "Adam: A method for stochastic optimization," 2014.

[30] A. Othmani, D. Kadoch, K. Bentounes, E. Rejaibi, R. Alfred, and A. Hadid, "Towards robust deep neural networks for affect and depression recognition from speech," in *International Conference on Pattern Recognition*, 2021, pp. 5-19: Springer .

[31] L. Van der Maaten and G. J. J. o. m. l. r. Hinton, "Visualizing data using t-SNE," vol. 9, no. 11, 2008.