INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

JAMFN: Joint Attention Multi-Scale Fusion Network for Depression Detection

*Li Zhou, Zhenyu Liu**, *Zixuan Shangguan, Xiaoyan Yuan, Yutong Li, and Bin Hu**

Lanzhou University, China

{zhoul2020, liuzhenyu, shanggzx20, yuanxy20, liyt20, bh}@lzu.edu.cn

Abstract

Recently, with the widespread popularity of the Internet, social networks have become an indispensable part of people's lives. As social networks contain information about users' daily moods and states, their development provides a new avenue for detecting depression. Although most current approaches focus on the fusion of multimodal features, the importance of fine-grained behavioral information is ignored. In this paper, we propose the Joint Attention Multi-Scale Fusion Network (JAMFN), a model that reflects the multiscale behavioral information of depression and leverages the proposed Joint Attention Fusion (JAF) module to extract the temporal importance of multiple modalities to guide the fusion of multiscale modal pairs. Our experiment is conducted on D-vlog dataset, and the experimental results demonstrate that the proposed JAMFN model outperforms all the benchmark models, indicating that our proposed JAMFN model can effectively mine the potential depressive behavior.

Index Terms: Depression detection, Vlog, Joint Attention Multi-Scale Fusion Network (JAMFN)

1. Introduction

Depression is a common mental illness that affects hundreds of millions of people of all ages worldwide [1]. According to a study by the World Health Organization, depression is currently ranked as the 3rd most significant economic burden of disease and is expected to grow, becoming the most serious disease by 2030 [2].

Depression is usually accompanied by cognitive, physical, and behavioral symptoms such as low interest in everything, depressed mood, waning energy, poor self-identity, and poor concentration, and untreated depressed patients may even have suicidal thoughts of self-harm [3]. The current clinical diagnosis relies mainly on the clinical experience of physicians. The diagnosis of depression by interviewing the patient has a high rate of misdiagnosis [4]. The reason is that most of the patients have a restrained mentality towards the doctors, which prevents them from communicating effectively with the patients. Therefore, it is of great significance to explore an objective and high accuracy auxiliary diagnostic method. Currently, auxiliary diagnosis of depression can be divided into two types based on physiological signals and non-physiological signals. Physiological signals include galvanic skin response [5], heart rate [6], electroencephalogram [7], and nuclear magnetic [8]. More researchers have developed studies on non-physiological signals such as speech [9], text [10] and facial expressions [11]. A study by Mehrabian et al [12] found that the amount of emotional information conveyed through facial expressions, acoustic features and text accounted for 55%, 38% and 7% of the total information, respectively. Therefore, the auxiliary diagnosis of depression based on visual and acoustic features has become a hot research topic.

Our contributions. To mine and fuse multimodal fine-grained features, we propose the Joint Attention Multi-Scale Fusion Network (JAMFN), a model for vlog-based depression detection with acoustic and visual features. The JAMFN model is designed in four steps: to begin with, BILSTM is used to extract the context-dependent high-level semantic features of each modality. Then, a series of convolutions are further used to generate multi-scale semantic features based on context-dependent features. Next, the proposed Joint Attention Fusion (JAF) module effectively leverages the temporal importance of multiple modalities to guide the same scale feature fusion of different modalities. In the end, the logical value of the depression detection label is inferred from the depression detection layer. Empirical results on D-vlog dataset demonstrate that the proposed JAMFN model outperforms all the benchmark models.

2. Related work

Shen et al [13] extracted six depression-related features from Twitter postings and used these features as the basis for constructing a multimodal learning dictionary for depression recognition, and they found that depressed users had nearly 200% more first-person pronouns in their postings compared to healthy people. Gui et al [14] proposed a cooperative multimodal approach for automatic selection of relevant images and texts, and experimental results showed that the model has good robustness. Chiong et al [15] proposed text preprocessing and text-based characterization methods for depression detection and demonstrated the generality of their methods through experiments across databases. Mann [16] found that a social media-based depression detection task can be modeled as a multiple-instance learning (MIL) problem that models user depressive behavior by mining the temporal dependencies between user posts. Cheng et al [17] used the user's posted text, image and time as input features, and then analyzed the importance of each user's post for depression detection using the T-LSTM model. Yoon et al [18] collected vlogs related to depression detection from YouTube, and thus constructed the D-Vlog dataset, and further proposed the Depression Detector model for depression detection on vlogs. Li et al. [19] proposed the TAMFN model to mine temporal information from vlog from unimodal and multimodal perspectives with good results.

3. Proposed Approach

Fig. 1 illustrates the overall depression detection framework proposed in this paper, which consists of two major components

10.21437/Interspeech.2023-183

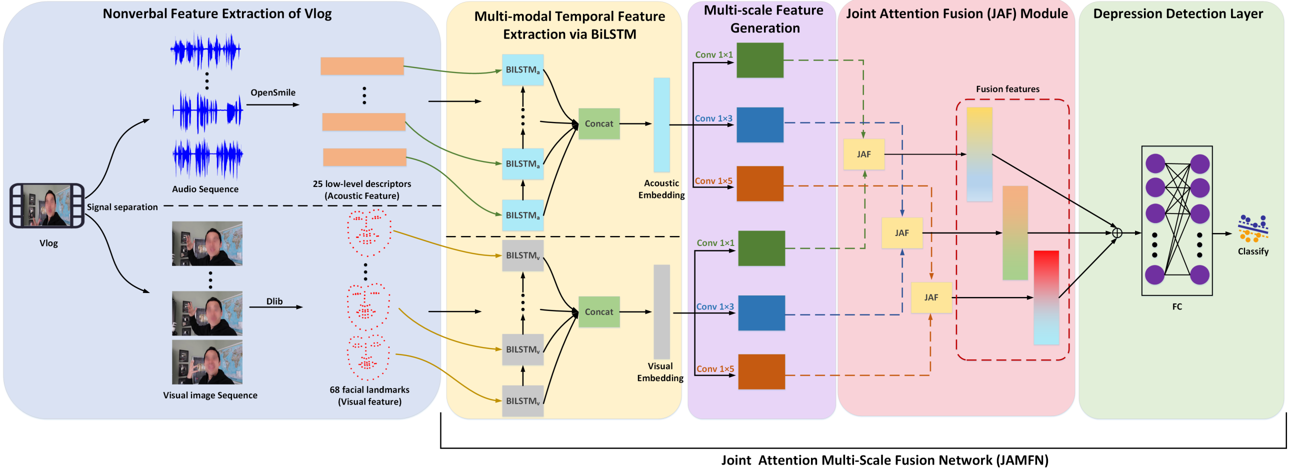
Nonverbal Feature Extraction of Vlog
Multi-modal Temporal Feature Extraction via BiLSTM
Multi-scale Feature Generation
Joint Attention Fusion (JAF) Module
Depression Detection Layer
OpenSmile
Audio Sequence
25 low-level descriptors (Acoustic Feature)
Signal separation
Vlog
Dlib
Visual image Sequence
68 facial landmarks (Visual feature)
BILSTMa
BILSTMv
Concat
Acoustic Embedding
Visual Embedding
Conv 1×1
Conv 1×3
Conv 1×5
JAF
Fusion features
FC
Classify
Joint Attention Multi-Scale Fusion Network (JAMFN)

Figure 1: *An illustration of the proposed overall depression detection framework. The first part of the framework is feature extraction of the vlog. Dlib toolkit [20] is used to extract facial landmarks of images, and OpenSmile toolkit [21] is used to extract low-level Descriptors of audio. The rest of the framework is the JAMFN model.*

for the extraction of acoustic and visual features of vlog and the JAMFN model for depression detection. Next, we illustrate each step in the JAMFN model in detail.

3.1. Multi-modal Temporal Feature Extraction via BiLSTM

Bidirectional LSTM (BiLSTM) [22] uses two LSTM models with different directions to learn the semantic dependency from front to back and from back to front respectively. Each layer of BiLSTM is composed of two LSTM models with opposite directions. The output of each layer of BiLSTM is shown as follows:

$$h_p^{(i)} = LSTM_{FW}(O_p^{(i-1)}) \tag{1}$$

$$H_{t-p}^{(i)} = LSTM_{BW}(O_{t-p}^{(i-1)}) \tag{2}$$

$$O_p^{(i)} = [h_p^{(i)}, H_{t-p}^{(i)}] \tag{3}$$

Where t is the length of the sequence, $LSTM_{FW}$ is the LSTM model learning forward semantic dependence, and $h_p^{(i)}$ is the output of $LSTM_{FW}$ at i-th layer at time p. $LSTM_{BW}$ is the LSTM model learning backward semantic dependence, and $H_p^{(i)}$ is the output of $LSTM_{BW}$ at time $t-p$ of the i-th layer. $O_p^{(i)}$ denotes the output of i-th layer at time p.

In this paper, we employ two BiLSTM models, $BiLSTM_a$ and $BiLSTM_v$, to capture the bidirectional semantic dependencies of acoustic and visual features, respectively. For the last layer of BiLSTM, we simply add forward and backward output features, and then concatenate the output features at all moments of the last layer of BiLSTM, as shown below:

$$O_p^{(n)} = h_p^{(n)} + H_{t-p}^{(n)} \tag{4}$$

$$BiLSTM_{Output} = Concat(O_1^{(n)}, O_2^{(n)}, ..., O_t^{(n)}) \tag{5}$$

where n denotes the last layer of the BiLSTM and $BiLSTM_{Output}$ is the final output of BiLSTM.

3.2. Multi-scale Feature Generation

Based on BiLSTM to extract multimodal context-dependent semantic features, we introduce a series of convolutions that provide different receptive fields to further capture the multiscale behavioral and local contextual relationships of acoustic and visual features in vlog. Specifically, we employ one-dimensional convolutions with convolution kernel sizes of 1, 3 and 5 to extract temporal multiscale features of acoustic and visual, respectively. In addition, we set the padding parameters to 0, 1, and 2 for the convolution kernels of sizes 1, 3, and 5, respectively, to make the features of each scale have the same dimension.

$$F_m^i = Conv_{1\times i}(F_m) \quad i \in \{1, 3, 5\}, m \in \{A, V\} \tag{6}$$

Here, F_A and F_V are context-dependent semantic features for acoustics and visuals, respectively. F_A^1 denotes the feature obtained by extracting F_A with a one-dimensional convolutional kernel of size 1, and so on for other cases.

3.3. Joint Attention Fusion (JAF) Module

For a given video sequence, acoustic and visual features contribute differently to each segment. Since multiple modalities convey different value information, their complementary relationships need to be captured efficiently. To reliably combine these modalities, we propose the Joint Attention Fusion (JAF) module, which relies on a joint attention-based fusion mechanism to efficiently encode the information between modalities, with the detailed structure shown in Fig. 2. Specifically, through the previous subsection, different scales of acoustic and visual features can be obtained, and we can construct 3 modality pairs, i.e., the same scale between different modalities constitutes a modality pair. For the fusion of a modal pair, first, the feature dimensions of the two modal features $X_A \in \mathbb{R}^{L\times H}$ and $X_V \in \mathbb{R}^{L\times H}$ are converted to 1 using two fully connected layers, respectively, to obtain the temporal features $T_A \in \mathbb{R}^{L\times 1}$ and $T_V \in \mathbb{R}^{L\times 1}$ of the two modalities, i.e., (L, H) is converted to $(L, 1)$, where L is the sequence length, and H is the feature representation dimension.

$$T_A = FC_1(X_A) \tag{7}$$

$$T_V = FC_2(X_V) \tag{8}$$

Then, the temporal features of the two modalities are concatenated to construct a joint feature representation T_{Joint}, and then a fully connected layer captures the temporal importance

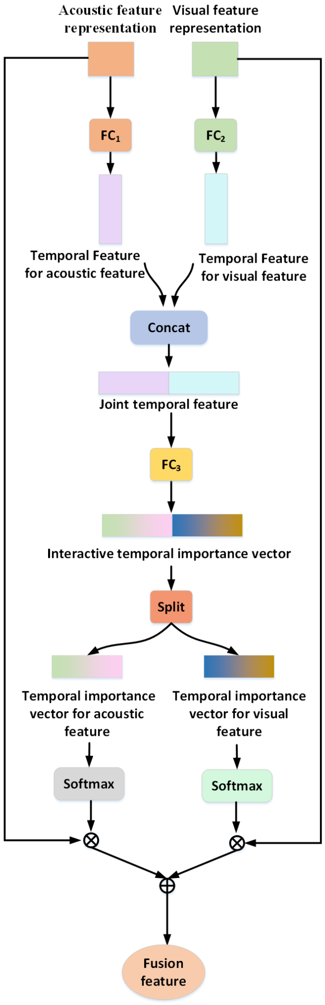
Acoustic feature representation
Visual feature representation
FC1
FC2
Temporal Feature for acoustic feature
Temporal Feature for visual feature
Concat
Joint temporal feature
FC3
Interactive temporal importance vector
Split
Temporal importance vector for acoustic feature
Temporal importance vector for visual feature
Softmax
Softmax
Fusion feature

Figure 2: *An illustration of the proposed JAF module.*

across the interaction between acoustics and visuals to obtain the interactive temporal importance vector $T_{Mixed} \in \mathbb{R}^{2L}$.

$$T_{Joint} = Concat(T_A, T_V) \tag{9}$$

$$T_{Mixed} = FC_3(T_{Joint}) \tag{10}$$

Finally, the interactive temporal importance vector T_{Mixed} is separated to obtain the temporal importance vectors $T_{A\_mixed} \in \mathbb{R}^L$ and $T_{V\_mixed} \in \mathbb{R}^L$ for each modality, and the temporal attention vectors for each modality are obtained by normalizing the two temporal importance vectors with the Softmax function, then the ultimate fusion feature $Fusion$ is given below.

$$T_{A\_mixed}, T_{V\_mixed} = Split(T_{Mixed}) \tag{11}$$

$$Fusion = Softmax(T_{A\_mixed}) \otimes X_A + Softmax(T_{V\_mixed}) \otimes X_V \tag{12}$$

Here, $\otimes$ is element-wise product.

3.4. Depression Detection Layer

With the JAF module, we can obtain the fusion features of the 3 modal pairs, and we simply add the fusion features at different scales. The reasoning of the ultimate logical value used for depression detection is as follows:

$$F\_out = Fusion^1 + Fusion^3 + Fusion^5 \tag{13}$$

$$\hat{Y} = FC(F\_out) \tag{14}$$

Here, $Fusion^1$, $Fusion^3$ and $Fusion^5$ denote the fusion features with scales 1, 3 and 5, respectively, and $\hat{Y}$ is the logical value. In this work, we utilize the cross-entropy loss function as the loss function for depression detection.

4. EXPERIMENTS

4.1. Dataset Introduction

The vlog data used in this paper comes from the D-vlog dataset constructed by Yoon et al [18], who analyzed videos posted on YouTube between January 1, 2020 and January 31, 2021 to collect the required vlogs based on keywords. the keywords for depression vlogs were ' depression daily vlog', 'depression journey', 'depression vlog', 'depression episode vlog', 'depression video diary', 'my depression diary ', 'my depression story', etc. The keywords for non-depression vlogs are 'daily vlog', 'grwm (get ready with me) vlog', 'haul vlog ', 'how to vlog', 'day of vlog', 'talking vlog' and so on. After a series of data cleaning and filtering operations, they finally obtained 961 labeled vlog data from 816 different people, with the number of depressed and non-depressed people being 555 and 406 respectively. The number of training set, validation set and test set of the D-Vlog dataset were 647, 102 and 212 respectively, i.e. the allocation ratio was 7:1:2.

4.2. Experimental Setup

In this paper, pytorch framework [23] is adopted to implement our model. The training, validation and testing operations of the model are carried out on NVIDIA PCIE A100 graphics card with 40G memory. The Adam optimizer [24] is used to optimize the weight update of the model, and the batch size, epoch, learning rate, weight decay, and eps of the Adam optimizer are set to 32, 30, 1e-4, 5e-4, and 1e-8, respectively. In the BiLSTM models for acoustic and visual modalities, we implement a 6-layer bidirectional with 200-dimensional hidden states. And the number of output channels of the multiscale convolution kernel is 512. In addition, we leverage the early stop mechanism to train the model to avoid overfitting, and the patience is set to 4. For the testing phase, the weighted average precision, recall, and f1 score are used to comprehensively evaluate the performance of the models.

4.3. Performance Evaluation on D-vlog Dataset

To verify the validity of our proposed JAMFN model, the performance of the JAMFN model is compared with the 10 benchmark models proposed by Yoon et al [18]. Table 1 reports the performance of the JAMFN model and other benchmark models on the D-vlog dataset. The weighted average precision, recall, and f1 of the JAMFN model reach 68.18 ($\times 10^{-2}$), 68.39 ($\times 10^{-2}$), and 68.25 ($\times 10^{-2}$), respectively. In comparing the experimental results, traditional machine learning methods (e.g., LR, SVM, and RF) performe unsatisfactorily due to their poor ability to fit nonlinear data, and JAMFN naturally outperformed the traditional machine learning models easily. Although deep learning models (e.g., BiLSTM, TFN, and Depression Detector) all outperform traditional machine learning models, their depression detection performance is also inferior to the JAMFN model. Further, the TAMFN model proposed by Li et

Table 1: *Performance of our approach with all benchmark models on the D-vlog dataset.*

Model	Precision	Recall	F1-Score
LR	54.86	54.72	54.78
SVM	53.10	55.19	52.97
RF	57.69	58.49	57.84
KNN-Fusion	57.86	59.43	54.25
BLSTM	60.81	61.79	59.70
TFN	61.39	62.26	61.00
Fusion_Concat	62.51	63.21	61.10
Fusion_Add	59.11	60.38	58.11
Fusion_Multiply	63.48	64.15	63.09
Depression Detector [18]	65.40	65.57	63.50
TAMFN [19]	66.02	66.50	65.82
JAMFN (proposed)	**68.18**	**68.39**	**68.25**

Table 2: *Importance analysis of multi-scale generation (MG) module. '-' is the removal of the multiscale generation module from the JAMFN model, while ✓ indicates the reservation of the module.*

MG Module	Precision ($\times 10^{-2}$)	Recall ($\times 10^{-2}$)	F1-Score ($\times 10^{-2}$)
-	62.73	63.20	62.83
✓	**68.18**	**68.39**	**68.25**

al. outperformed all benchmark models, while our model similarly outperformed the TAMFN model, with our weighted average of precision, recall, and f1 improving by 3.2%, 2.8%, and 3.6%, respectively, over the TAMFN model. From the comparison experiments, it is demonstrated that our proposed JAMFN model achieves the best performance in depression detection, indicating that the JAMFN model can mine more behavioral information related to depression.

4.4. Ablation Studies

In this subsection, we first explore the effect of different combinations of convolutional kernels, then, analyze the effectiveness of the JAF module, and finally, investigate the effect of leveraging different BackBones in the JAMFN model on the model performance.

Study on multi-scale generation module. In Table 2, we compare the model performance of removing and retaining the multi-scale generation (MG) module in the JAMFN model to explore the importance of the MG module. The experimental results show that the model with the MG module removed performs relatively poorly, the reason is that it lacks the fine-grained multi-scale features provided by the MG module, which proves the effectiveness of the MG module.

Study on the effect of convolution kernel size in MG module. We perform ablation experiments on multi-scale generation (MG) module to explore the effect of convolution kernel size. The results in Table 3 report that when the MG module lacks convolutional kernels of size 5, it makes the model perform the worst, which indicates that a large size convolutional kernel is more important. That is, a larger convolutional kernel has a larger field of perception and can capture more global features.

Study on JAF module. To justify our design of JAF module, we perform ablation experiments on feature fusion. Specifically, we take the JAF module and compare it with Add [25], Multiply [26], and Concat [27], three common feature fusion operations. More precisely, we simply replace the JAF module in the JAMFN model with Add, Multiply and Concat for feature fusion operations, respectively. Table 4 reports the performance of the different feature fusion approaches, which demonstrates that using the JAF module outperforms the other feature fusion operations, indicating that the JAF module can encode fused features more efficiently.

Table 3: *The Effect of Convolution Kernel Size in MG Module.*

Module	Precision ($\times 10^{-2}$)	Recall ($\times 10^{-2}$)	F1-Score ($\times 10^{-2}$)
w/o k=1	64.53	65.09	64.43
w/o k=3	66.02	66.50	65.82
w/o k=5	60.18	60.84	60.28

Table 4: *Comparisons between Different Feature Fusion Operations.*

Operation	Precision ($\times 10^{-2}$)	Recall ($\times 10^{-2}$)	F1-Score ($\times 10^{-2}$)
Add	63.42	59.43	59.21
Multiply	61.09	61.32	61.18
Concat	65.22	63.20	63.40
JAF	**68.18**	**68.39**	**68.25**

5. Conclusions

In this paper, we propose the Joint Attention Multi-Scale Fusion Network (JAMFN) for vlog-based depression detection. The JAMFN model mines the fine-grained behavior of multiple modalities through the multiscale generation module, while the JAF module is used to extract the temporal importance between modalities to guide the fusion of modal pairs at multiple scales. We evaluate the performance of the proposed model on the D-vlog dataset, and the experimental results show that the proposed method is effective. Moreover, the results of the ablation experiments illustrate that the multi-scale generation (MG) module and the JAF module have a positive impact on the depression detection ability of the JAMFN model.

Although the JAMFN model achieves the best performance on the D-Vlog dataset, there is still much room for improvement here. Since the D-Vlog dataset is subjectively labeled by humans, it is inevitable that mislabeling will occur, and these mislabeling samples may firstly, overfit the model and secondly, affect the model to learn effective features. Therefore, in future work, we will focus on how to suppress the influence of the mislabeling samples in the training samples and thus improve the learning efficiency of the model.

6. Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2019YFA0706200), in part by the National Natural Science Foundation of China (Grant No.61632014, No.61627808, No. 61802159, No. 61802158). in part by Fundamental Research Funds for Central Universities (lzujbky-2019-26, lzujbky-2021-kb26).

7. References

[1] Moustafa, A. Ahmed, Misiak, Blaej, Tindle, Richard, Frydecka, and Dorota, "Impulsivity and its relationship with anxiety, depression and stress," *The Nature of Depression*, pp. 183–194, 2021.

[2] G. S. Malhi and J. J. Mann, "Depression." *The Lancet*, vol. 392, 2019.

[3] K. Hawton, C. C. i Comabella, C. Haw, and K. Saunders, "Risk factors for suicide in individuals with depression: a systematic review," *Journal of affective disorders*, vol. 147, no. 1-3, pp. 17–28, 2013.

[4] A. J. Mitchell, "Clinical utility of screening for clinical depression and bipolar disorder," *Current opinion in psychiatry*, vol. 25, no. 1, pp. 24–31, 2012.

[5] M. Naszariahi, K. N. haleeda, and N. A. M. Mortar, "The development of galvanic skin response for depressed people," in *AIP Conference Proceedings*, vol. 2291, no. 1. AIP Publishing LLC, 2020, p. 020096.

[6] D. Kuang, R. Yang, X. Chen, G. Lao, F. Wu, X. Huang, R. Lv, L. Zhang, C. Song, and S. Ou, "Depression recognition according to heart rate variability using bayesian networks," *Journal of psychiatric research*, vol. 95, pp. 282–287, 2017.

[7] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level fusion approaches based on multimodal eeg data for depression recognition," *Information Fusion*, vol. 59, pp. 127–138, 2020.

[8] J. YAN, F. ZHANG, M. WANG, H. TANG, Q. HU, and X. ZHANG, "Classification of rock-electro parameters of low-permeability sandstone based on nuclear magnetic resonance log and its appication: An example of e s4 in south slope of the dongying depression," *Chinese Journal of Geophysics*, vol. 62, no. 7, pp. 2748–2758, 2019.

[9] 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.

[10] K. Sampath and T. Durairaj, "Data set creation and empirical analysis for detecting signs of depression from social media postings," in *International Conference on Computational Intelligence in Data Science*. Springer, 2022, pp. 136–151.

[11] W. Guo, H. Yang, Z. Liu, Y. Xu, and B. Hu, "Deep neural networks for depression recognition based on 2d and 3d facial expressions under emotional stimulus tasks," *Frontiers in Neuroscience*, vol. 15, p. 609760, 2021.

[12] A. Mehrabian and J. A. Russell, *An approach to environmental psychology.* the MIT Press, 1974.

[13] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution." in *IJCAI*, 2017, pp. 3838–3844.

[14] 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, no. 01, 2019, pp. 110–117.

[15] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Computers in Biology and Medicine*, vol. 135, p. 104499, 2021.

[16] P. Mann, A. Paes, and E. H. Matsushima, "Screening for depressed individuals by using multimodal social media data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 18, 2021, pp. 15 722–15 723.

[17] J. C. Cheng and A. L. Chen, "Multimodal time-aware attention networks for depression detection," *Journal of Intelligent Information Systems*, pp. 1–21, 2022.

[18] 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*, 2022.

[19] L. Zhou, Z. Liu, Z. Shangguan, X. Yuan, Y. Li, and B. Hu, "Tamfn: Time-aware attention multimodal fusion network for depression detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.

[20] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "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, 2015.

[22] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[23] E. Stevens, L. Antiga, and T. Viehmann, *Deep learning with PyTorch.* Manning Publications, 2020.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human-computer interaction*. Springer, 2008, pp. 92–103.

[26] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 02, 2020, pp. 1359–1367.

[27] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 433–436.