International Journal of Engineering Sciences 2022 15(1) 24 - 32

*INTERNATIONAL JOURNAL*
*OF*
*ENGINEERING SCIENCES*

# DepressNet: A Multimodal Hierarchical Attention Mechanism approach for Depression Detection

Guramritpal Singh Saggu*, Keshav Gupta, K. V. Arya
Multimedia and Information Security Research Group,
ABV-Indian Institute of Information Technology and Management, Gwalior, India
Ciro Rodriguez Rodriguez[@]
[@]Department of Software Engineering, Universidad Nacional Mayor de San Marcos,
Lima, Peru

(Received 30 November 2021; Accepted 12 March 2022)

DOI: https://doi.org/10.36224/ijes.150104

**Abstract**
Depression has been the prime cause of mental-health illness globally. A major depressive disorder is a common mental health disorder that affects both psychologically and physically, which could lead to the loss of life in extreme cases. Detection of depression from the recording of an interview could help with early diagnosis. This paper proposes a three-stage framework multimodal machine learning approach called DepressNet for depression detection using the PHQ-8 Questionnaire score. Bidirectional Long Short-Term Memory (BLSTM) layer network has been proposed, and Extended Distress Analysis Interview Corpus (E-DAIC) dataset was used for the training and validation of the proposed method with the uses multiscale temporal features from audio, video, and text modality and attention mechanisms for fusion. The method achieved the RMSE of 4.32 and CCC of 0.662 on the development set, and on the test set, we got RMSE of 5.36 and CCC of 0.457, outperforming the other methods.

**Keywords**: Depression, multimodal, attention mechanism, multiscale temporal, audio-visual-text, BLSTM

## 1. Introduction

A person's emotions and behaviour show a great impact on an individual's physical and mental health. Mental Illness includes issues such as depression, bipolar disorder, an unhealthy state of mind, spectrum disorder, and many more. Mental illness is one of the most widespread and common health problems among the public. This fact is supported by the numbers that around 450 million people suffer from such issues and problems worldwide [1]. Not only adults but children and adolescents under the age of 18 years are also facing issues of mental health disorders and illness. Depression is one of the leading causes of such disorders and thus resulting in an increased possibility of suicide attempts [2]. As such mental health conditions may prove to be fatal, thus early detection becomes more vital so that it may help to understand the health condition in a better way and provide the patient with better care [3]. Diagnosis of such mental health disorders is based on self-report and related health questionnaires, which are specifically designed to detect certain feelings or interactions rather than relying on any lab test. As the data available for such studies related to mental health are increasing exponentially [4], the use of artificial intelligence and machine learning technologies comes into play, and such technologies are increasingly used to help mental health providers better understand the mental conditions, thus helping in improved decision making. Deep learning is one of the latest advances in this field of artificial intelligence which mimics the workings of the human brain in processing the data and the

*Corresponding author
Email address: imt_2017036@iiitm.ac.in (Guramritpal Singh Saggu)

ISSN 0976 – 6693.

ability to learn, recognize and detect the patterns and thus helping in decision making. Deep learning algorithms are able to learn without human supervision and have shown better performance in many data-driven applications in various sectors. Deep learning [5] has vast application in healthcare and has already been extensively used for medical image analysis, pattern detection and recognition to identify various fatal and rare diseases. Thus, deep learning has come a long way in healthcare [6].

Major depressive disorder, or simply depression [7], is a state in which an individual suffers from low mood and has a lack of motivation to do any kind of activity. It has a negative impact on a person's thoughts, feelings, and motivation. In the worst of cases, depression can lead a person to death as well. WHO has estimated that more than 300 million people are suffering from depression, which makes it a very serious problem for society. Since the current depression diagnosis is dependent on clinical interviews conducted by psychiatrists [8], it is subjective, inefficient, and consumes a lot of expert labour. Clinicians use self-report questionnaires to track the symptoms and their seriousness to diagnose or check the progress of depression. Some commonly used self-report questionnaires are the Hamilton Depression Rating Scale, Patient Health Questionnaire, and Beck's Depression Inventory. This method of diagnosis or tracking is not very reliable since it is very much dependent on the emotional state of a person.

Another problem is that this way makes it impossible to distinguish between a person who complains a lot and a person suffering from moderate or mild depression. Also, it has been observed that individuals suffering from severe depression do not talk much. These unreliability issues and subjectivity make the accurate early diagnosis of depression very challenging, and many people are misdiagnosed. Furthermore, since depression is not a physical ailment, there is no straightforward diagnostic test for the same, and the only way is to interview the individual routinely. As it has been estimated, depression affects more than 300 million people globally and affects the motivation of the person. In addition, depression has a huge impact on the economy. It has been estimated that there is a loss of more than 210 Billion USD because of depression [9]. So the quest of finding an effective diagnosis for depression has become a top priority in recent times.

The major contribution in this paper are listed below:

1. Depression detection network using Attention-based fusion of modalities: We present an attention layer-based fusion network to perform mid fusion audio, visual, and text and understand the importance of each for depression detection.
2. Fusion of low-level descriptors and Deep Representations: We used multiple low-level descriptors and CNN based representations of modalities and performed mid fusion on them.

## 2. Literature review

Depression detection and its analysis use multimodal inputs, and here we discuss the work done in different modalities such as audio, visual, text, and multimodal sentiment analysis. In [10], authors have studied the relation between distress and existing facial emotion recognition through meta-analysis. Poria et al. [11] counted on sequence modeling and temporal long short term memory (LSTM) methods for capturing contextual information from visual in sentiment analysis tasks. To estimate the head pose movements and the action units against them Valstart et al. [12] introduced the FERA dataset, which quantifies facial expressions in challenging scenarios. Finally, in [13], Emotion Recognition in the Wild (EmotiW) Challenge dataset was introduced and utilized a hybrid convolutional neural network-recurrent neural network (CNN-RNN) framework for facial expression analysis. These datasets have been very crucial in advancing state-of-art research around facial expression recognition and distress prediction.

Baltrušaitis et al. [14] introduced an open-source interactive tool to estimate facial behaviour, OpenFace is a widely used tool used in this work. OpenFace is used to extract low-level descriptors from visual modality and face land-mark regions. Studying the multimodal approaches that have been used for distress detection. Lamet al. [15] uses context-aware feature creation methods to analyze the level of depression. They also used the end-to-end trainable deep neural networks to analyze the levels.
Furthermore, they introduced the infusion of data augmentation techniques based on topic modeling in the transformer network. Morales et al. [16] studied various fusion techniques and provided their review for the task of detecting depression. They also introduced a new fusion approach for depression detection, which is based on computational linguistics. For human behaviour analysis, in [17], authors released a multimodal spontaneous emotion corpus. Poria et al. [18] harvest emotions in multimedia content using the fusion of video, audio, and text. Feature and decision level fusion techniques were used in order to perform the fusion. Alghowinem et al. [19] perform multi-modal depression detection with the help of eye gaze and head poses fixation. The classification of either depressed or healthy is done with the help of statistical tests on the selected features. We have used some advanced sequence modeling approaches and built our solution around that.

## 3. Problem formulation

The problem at hand is to detect depression from the interview videos of the subjects answering the PHQ-8 questionnaire. The proposed method for depression detection uses audio, visual, and text modalities. For visual modality, we have low-level descriptors, facial action units, pose and gaze descriptors. We have used extended Gneva minimalist acoustic parameter set (eGeMAPS) and Mel frequency cepstral coefficients (MFCCs) for audio. Universal sentence encodings have been used for text modality. Figure 1 shows the extraction and flow of data from the database to the model.

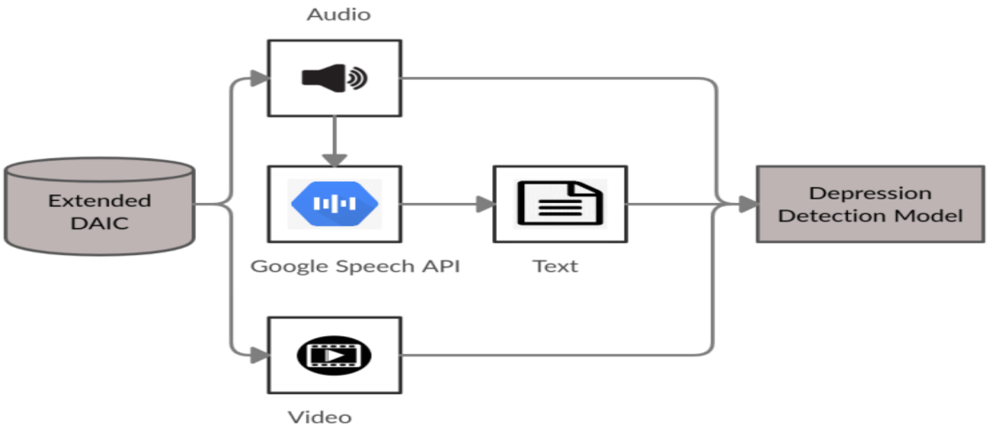


Figure 1: High-level view of the depression detection model

The proposed method divides the model into two parts: the intramodality part and the intermodality part. In the intra modality, each modality is processed in isolation, and only fusion of features of a single modality occurs, and in intermodality, the processed features from the modalities are fused together and processed further in order to obtain results.
*Intra modality Hierarchy:*
For audio and visual low-level descriptors, we first extract multi-scale temporal features using an inception module; extracted features are then passed through a 200 units Bidirectional Long Short Term

(BLSTM) network to identify patterns in the temporal features, the output of the BLSTM network for low-level descriptors is fused with other outputs by concatenating and passing thought an attention layer so that each modality would have a single stream of output. For text, the universal sentence encodings are passed through a 200 unit BLSTM network.

*Inter modality Hierarchy:*

Intra modality network is trained for audio, visual, and text, and the output of each modality is concatenated and fused using the attention mechanism. The output obtained from this attention layer is forwarded to a 200 unit BLSTM network, which is followed by two dense connected layers which produce the output in terms of Patient Health Questionnaire score (PHQ-8)

## 4. Method of analysis

The architecture developed in the work is shown in Figure 2.

### 4.1 Feature Extraction

Feature extraction is the process of extraction of particular information from an initial set of raw data. The method used extracts low-level descriptors from the audio and visual data. Low-level descriptors come from the expert knowledge of the signal. Low-level descriptors aim to describe the signal at every instant of time. In the case of audio descriptors include spectral, prosodic, and voice quality information, and for video, they include appearance and geometric information. Low-level descriptors being used might depend on the task at hand. Open source tools like OpenFace and OpenSmile are used to extract low-level descriptors from video and audio, respectively.

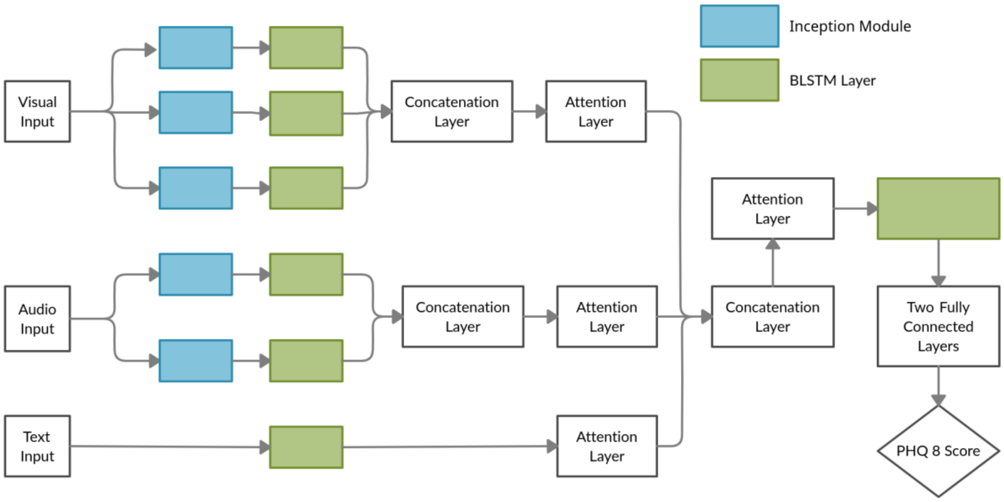


Figure 2: DepressNet: Multimodal Hierarchical Attention mechanism-based network

OpenFace, used to extract visual features, gives 17 FAUs for every visual frame with confidence. Additionally, pose and gaze descriptors are extracted. The audio channel has the information of voice quality, cepstral, and prosodic, whereas the visual has the geometric and appearance details. For audio features, we have used eGeMAPS, as stated in section 3, which has 88 measures that cover the acoustic dimensions. OpenSmile toolkit is used to compute the acoustic LLDs, which use MFCCs with their delta and double-deltas i.e. first-order and second-order derivatives. eGeMAPS are computed over every window, and all other LLDs are computed in the time domain by using mean and standard deviation over a sliding window of 4 seconds length and a hop size of 1 second.

### 4.2 Multiscale Temporal Features Inception Module

We use the inception module to extract the multi-scale temporal information from the initial low-level descriptors. The Inception module consists of different kernel sizes, each kernel convolves over the low-level descriptors to provide processed temporal information, and since we use multiple different kernel sizes, we get different scales of processed temporal information. The multi-scale way would help us represent the complex dynamic information effectively, hence better recognizing the characteristics of depression.

### 4.3 Attention Mechanism

A sequence model has two components, an encoder, and a decoder. The context vector is responsible for encoding the information, but they could not preserve the information about subjects and entities. Therefore, the concept of attention was introduced in which, at each decoding step, the decoder gets to look at any particular state of the encoder and thus becomes a weighted sum of all the previous encoder states. So, for example, we use the attention mechanism for the fusion task, and we first concatenate all the tensors that we have to fuse; the final concatenated tensor is then passed through the attention layer, which outputs the weighted tensor in the context of the task in hand.

In our method, the attention mechanism is primarily used for fusion. Tensors to be fused are first concatenated with the help of a concatenation layer and then passed through the attention layer, which outputs the weighted tensor in the context of the task at hand. The architecture invokes this fusion mechanism twice, the first time while combining the features from different low-level descriptors of the same modality and a second time while combining the features from different modalities.

### 4.4 Sparse & Bidirectional Long Short Term Memory Network

Bidirectional LSTM(BLSTM) trains two recurrent networks; one is trained on the input sequence and another on a time-reversed copy of the input sequence, as shown in Fig. 2. This two-way training helps network with additional context and can help to give better results than recurrent network structures. LSTM was proposed in [12], and to capture and preserve information over a long history, it uses a gate mechanism containing an input gate $i$t, an output gate $o$t,d a forget gate $f$t, and a hidden state $h$t. The transition equations are given as follows for the sequence of inputs {x1, x2,....xT}

$$i_t = \sigma(W^{xi}x_t + W^{hi}h_{t-1} + b_i)$$

$$f_t = \sigma(W^{xf}x_t + W^{hf}h_{t-1} + b_f)$$

$$o_t = \sigma(W^{xo}x_t + W^{ho}h_{t-1} + b_o)$$

$$u_t = \sigma(W^{xu}x_t + W^{hu}h_{t-1} + b_u)$$

$$c_t = i_t \odot u_t + f_t \odot c_{t-1}$$

$$h_t = o_t \odot tanh(c_t)$$

Where and $\sigma(x) = \frac{1}{1+exp(-x)}$.

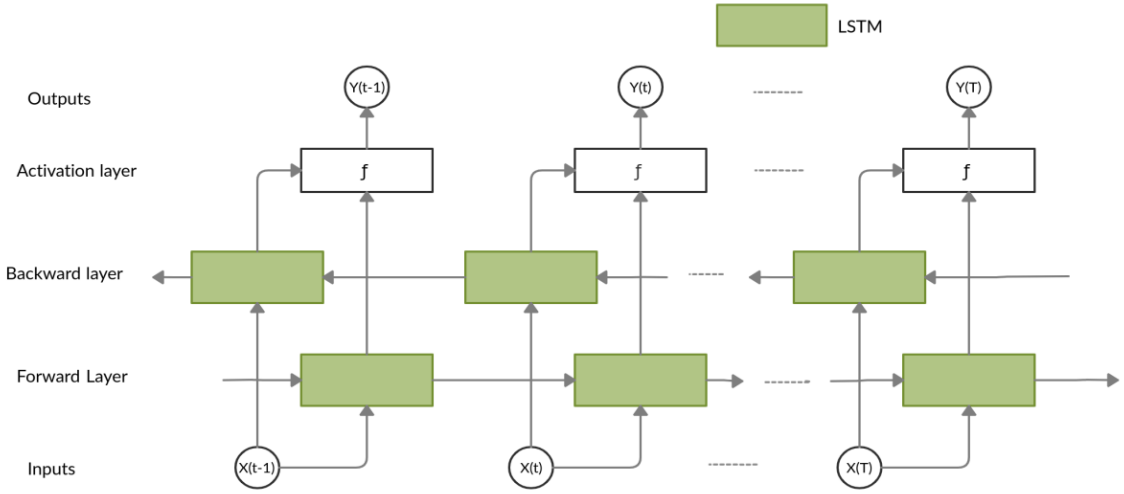


Figure 3: Architecture of Bidirectional Long Short Term Memory Unit

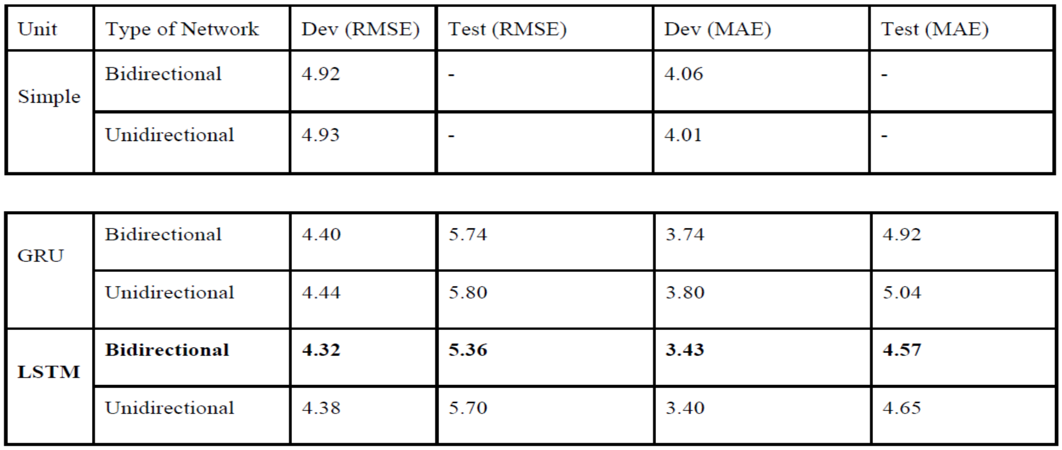
## 5. Results and discussion

Extended Distress Analysis Interview Corpus(E-DAIC), a private dataset, has been used in our study. This dataset is an expansion of the distress analysis interview corpus, which came out earlier in 2018. The dataset contains audio and video recordings of clinical interviews for mental disorders such as depression. In the initial dataset, the interviews were performed by a bot controlled by a person in another room, but in the expansion, all the interviews for data collection were performed by an Artificial Intelligence bot only. The dataset has been split into train, development, and test with appropriate gender distribution. From the total of 275 sessions, a number of sessions are split into 163, 56, and 56 for train, development, and test, respectively.

In the task of depression detection, we had to predict the PHQ-8 score (Range 0-24) a person would get based on their interview. The performance of the proposed methodology is measured using the root mean squared error. Through our various experiments, we managed to achieve the best RMSE of 4.32 and CCC of 0.662 on the development set, and on the test set, we got RMSE of 5.36 and CCC of 0.457, which outperforms the result of the previous state-of-the-art models.

We have experimented with different recurrent units such as GRU and LSTM and their combination with unidirectional and bidirectional. As shown in Table 5.1, the results of different units depict that LSTM works best on both the development and test set. Bidirectional LSTM results are reported in Table 5.1, and they perform better because of their better memory retention ability as they use two more gates(Forget and Output) in addition to the update gate which is there in GRU, while there are no such gates in simple units. Thus, due to the more controlling ability of the LSTM, they give better results.

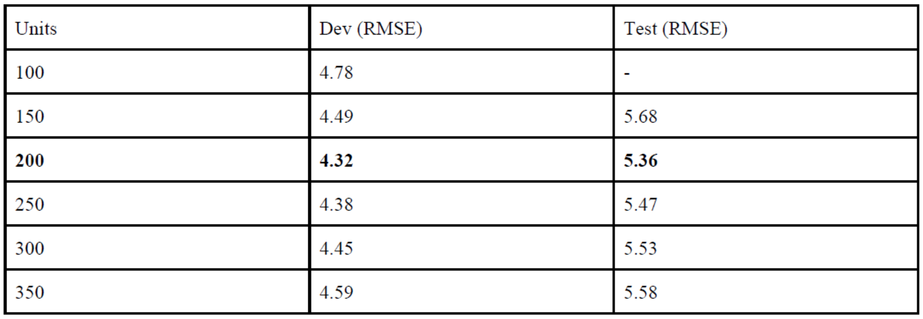
Table 5.1: Results of different types of RNN units on development and test set of E-DAIC dataset in terms of RMSE



| Unit | Type of Network | Dev (RMSE) | Test (RMSE) | Dev (MAE) | Test (MAE) |
|---|---|---|---|---|---|
| Simple | Bidirectional | 4.92 | - | 4.06 | - |
| | Unidirectional | 4.93 | - | 4.01 | - |
| GRU | Bidirectional | 4.40 | 5.74 | 3.74 | 4.92 |
| | Unidirectional | 4.44 | 5.80 | 3.80 | 5.04 |
| **LSTM** | **Bidirectional** | **4.32** | **5.36** | **3.43** | **4.57** |
| | Unidirectional | 4.38 | 5.70 | 3.40 | 4.65 |

We have experimented using different numbers of LSTM units in each layer. Having too few units could result in underfitting while increasing the number of units by too much could result in overfitting. Since we wanted to avoid both of them, we started from smaller units, increased them slowly, and tested the same. The end layers with 200 units performed best on the development set and test set.

Table 5.2: Results of using different numbers of units in RNN layer on development and test set of E-DAIC dataset in terms of RMSE.



| Units | Dev (RMSE) | Test (RMSE) |
|---|---|---|
| 100 | 4.78 | - |
| 150 | 4.49 | 5.68 |
| **200** | **4.32** | **5.36** |
| 250 | 4.38 | 5.47 |
| 300 | 4.45 | 5.53 |
| 350 | 4.59 | 5.58 |

Table 5.3: Results of the proposed method have been compared with previous models on development and test set on RMSE and CCC metrics.

| Model | Dev (RMSE) | Test (RMSE) | Dev (CCC) | Test (CCC) |
|---|---|---|---|---|
| Yin et al. [21] | 4.94 | 5.50 | 0.402 | 0.442 |
| Fan et al. [22] | 5.07 | 5.91 | 0.466 | 0.430 |
| Ringeval et al. [20] | 5.03 | 6.37 | 0.336 | 0.120 |
| Makiuchi et al. [23] | 3.86 | 6.11 | 0.696 | 0.403 |
| **Proposed Method** | **4.32** | **5.36** | **0.662** | **0.457** |

## 6. Conclusion

In this research work, we have worked on multimodal depression detection from audio, visual, and text modalities. A unique three-stage framework called DepressNet for depression using attention and BLSTM layer network has been proposed. The E-DAIC dataset is used for the training and validation of the proposed method. The proposed method designed using three modalities achieve the root means squared error of 4.32 on development and 5.36 on the test set. We achieved a better performance using a much more robust approach. In our method, we have used low-level descriptors of all the modalities. In future work, the improvement in the feature extraction module can be investigated. Additionally, more advanced fusion methods can be explored and worked on developing explainable models and the influence of features.

## References

[1] W. H. Organization, "The World Health Report 2001: Mental health, new understanding, new hope," World Health Organization, 2001.

[2] C. Su, Z. Xu, J. Pathak, and F. Wang, "Deep learning in mental health outcomes research: a scoping review," Translational Psychiatry, vol. 10, no. 1, pp. 1–26,2020.

[3] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," Annual review of clinical psychology, vol. 14, pp. 91–118, 2018.

[4] M. Hamilton, "Development of a rating scale for primary depressive illness," British journal of social and clinical psychology, vol. 6, no. 4, pp. 278–296, 1967.

[5] S. Vieira, W. H. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlate of psychiatric and neurological disorders: Methods and applications," Neuroscience & Biobehavioral Reviews, vol. 74, pp. 58–75, 2017.

[6] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities, and challenges," Briefings in bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018.

[7] R. Belmaker and G. Agam, "Major depressive disorder," New England Journal of Medicine, vol. 358, no. 1, pp. 55–68, 2008.

[8] W. Fan, Z. He, X. Xing, B. Cai, and W. Lu, "Multi-modality depression detection via multi-scale temporal dilated CNNs," in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC 2019), 2019, pp. 73–80.

[9] A. Morin, "Depression statistics everyone should know," available at URL: https://www.verywellmind.com/depression-statistics-everyone-should-know-4159056, 2019.

[10] M. Dalili, I. Penton-Voak, C. Harmer, and M. Munafò, "Meta-analysis of emotion recognition deficits in major depressive disorder," Psychological medicine, vol. 45, no. 6, pp. 1135–1144, 2015.

[11] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in Proceedings of the 55th annual meeting of the association for computational linguistics, 2017, pp. 873–883.

[12] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang,L. Yin, and M. Pantic, "FERA 2017-addressing head pose in the third facial expression recognition and analysis challenge," in Proceedings of 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017, pp. 839–847.

[13] C. Lu, W. Zheng, C. Li, C. Tang, S. Liu, S. Yan, and Y. Zong, "Multiple Spatio-temporal feature learning for video-based emotion recognition in the wild," in Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018, pp. 646–652.

[14] T. Baltrušaitis, P. Robinson, and L. -P. Morency, "Openface: an open-source facial behaviour analysis toolkit," in Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV 2016), 2016, pp. 1–10.

[15] G. Lam, H. Dongyan, and W. Lin, "Context-aware deep learning for multimodal depression detection," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2019), 2019, pp. 3946–3950.

[16] H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski and Lyle Ungar, "Towards Assessing Changes in Degree of Depression through Facebook," in Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, 2018.

[17] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin, "Multimodal spontaneous emotion corpus for human behaviour analysis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3438–3446.

[18] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," Neuro-computing, vol. 174, pp. 50–59, 2016.

[19] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviours," IEEE Transactions on Affective Computing, vol. 9, no. 4, pp. 478–490, 2016.

[20] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E. -M. Messner et al., "Workshop and Challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition," in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC 2019), 2019, pp. 3–12.

[21]. S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC 2019), 2019.

[22]. W. Fan, Z. He, X. Xing, B. Cai, and W. Lu, "Multi-modality Depression Detection via Multi-scale Temporal Dilated CNNs," in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC 2019), 2019.

[23]. M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of BERT-CNN and gated CNN representations for depression detection," in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC 2019), 2019.