Early adolescent depression detection system based on the transformer model

Haoze Yu*, Hongyu Shen, Jiarong Zhang, Zhengtao Liu
College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110169, China
* Corresponding author: 20227350@stu.neu.edu.cn

ABSTRACT

Depression is one of the more prevalent mental health disorders, characterized primarily by low mood. Recently, there has been a striking shift, and increasingly younger demographics have started showing symptoms of the same, particularly in adolescents. This paper proposes a new method for the automatic detection of incipient adolescent depression by use of deep multimodal learning techniques. This aims at improving preparedness to better face the rising problem of adolescent depression. In the proposed approach, unimodal features are extracted from electroencephalography(EEG), electrocardiogram(ECG), and speech signals using the Transformer model and subsequently fused into a comprehensive multimodal feature set for binary classification. The model does not only increase its generalizability by fusing different physiological signals but also increases the accuracy and reliability of diagnostic results by fusing multimodal features.

Keywords: Automatic detection of depression, deep learning, Transformer model, feature extraction, multimodal feature fusion

1. INTRODUCTION

Depression significantly impedes the physical and mental development of adolescents and stands as one of the most prevalent mental disorders [1]. Its primary symptoms include a persistent low mood, a lack of interest, sleep disruptions, and fatigue [2]. Contributory factors to adolescent depression include a familial history of the disorder and psychological stress stemming from social environments [3]. Annually, mental health disorders affect about 13% of children, 46% of adolescents, and 19% of adults worldwide [4], with a significant number of these cases attributable to adolescent depression. A report from 2017 estimated that depression afflicted approximately 322 million people globally [5]. During the pandemic in 2021, instances of severe depression doubled [6], indicating that depression could be affecting one in every ten individuals in our vicinity. Depression is characterized by its high prevalence, significant cure rates clinically, low acceptance of treatment, and high likelihood of recurrence. The hallmark of depression is a substantial and enduring low mood. Traditional diagnostic approaches, primarily reliant on the Beck Depression Inventory [7] or self-reports from patients along with the clinical acumen of physicians [8], currently dominate. Yet, these methods often yield low detection rates. Contributing factors include the pressures of contemporary social life, work and financial stresses, intense pressures from interpersonal relationships, limited societal acceptance of those with depression, and an overarching neglect of individuals suffering from the condition. Additionally, patients frequently lack the necessary skills to effectively manage stress and depressive symptoms, compounded by the limitations of current medical practices and the lack of highly effective diagnostic tools. In a society that is rapidly changing, the need to improve diagnosis and treatment of depression has brought into sharp focus increasing development across multifarious sectors.

Detection of depression is difficult, as the symptoms are subtle and often imperceptible to others, and the causes are complex and multifactorial, ruling out specialized laboratory testing. The clinically detected depression depends on the assessment of the case history—involving personal and family history—and measurement of a psychological state by voice, movement, and facial expressions [9]. This paper proposes a novel methodology of automatic detection of depression by adopting multimodal feature analysis. Following this, a multimodal approach has to be employed for the exact detection of depression, which is intricate and extremely multi-faceted.

There are inherent limitations in the current approaches for automatic early-stage depression detection. Brain reflection, psychological expression, and vocal expression of depression are so distinct that no single perspective can suffice in assessing it. On this regard, this paper uses a Transformer model to extract electroencephalography (EEG) signals,

International Conference on Future of Medicine and Biological Information Engineering (MBIE 2024), edited by Yudong Yao, Xiaoou Li, Xia Yu, Proc. of SPIE Vol. 13270, 132700T · © 2024 SPIE · 0277-786X · doi: 10.1117/12.3048175

electrocardiogram (ECG) signals, and speech signals into unimodal features and then combine all of them into a unified multimodal dataset. That is to say, the integration of these features depends on an automatic detection model that provides important technical support for computer-aided diagnosis, hence being much more convenient and effective as a diagnosis solution. This will enhance the diagnostic capability of clinicians and bridge the inadequacy of traditional methods.

2. RELATED WORK

2.1 Physiological signal-based detection techniques

Recent research has highlighted the use of physiological signals for depression detection. Methods focus on analyzing data from electroencephalography (EEG) and electrocardiography (ECG) to diagnose depression.

Cai et al. [10] developed a portable tri-electrode EEG system identifying depression through theta wave power. Their system showed that K-nearest neighbors (KNN) achieved the highest accuracy of 79.27% in classifying depression. Khandoker et al. [11] explored cardiac autonomic function, including heart rate variability and Tone-Entropy, to assess suicide risk in depressed patients. Significant differences were found in these measures between patients with and without suicidal ideation.

Despite advancements, physiological signal-based techniques face limitations. The reliability of depression diagnosis using EEG and ECG is variable, and high hardware costs and technical complexity restrict widespread clinical use.

2.2 Multimodal assessment methods based on behavior and language

Multimodal approaches, combining behavioral and linguistic data, offer promising methods for depression detection.

Ye et al. [12] proposed a deep learning-based method that integrates audio and text features. Their model, using Segmental Emotional Speech Experiment (SESE) and DeepSpectrum features, achieved an accuracy of 0.912 and an F1 score of 0.906. Gong and Poellabauer [13] emphasized the importance of multimodal data analysis for detecting depression, noting that Topic Models capture long-term information better than traditional methods. This approach provides deeper insights into depression's subtle expressions.

Lin et al. [14] introduced SenseMood, a system using CNN-based image analysis and BERT-based text analysis to detect depression from social media content. SenseMood offers detailed analysis reports useful for early detection and intervention. Deshpande and Rao [15] used emotional AI to analyze tweets for depressive content. Their multi-class Naive Bayes classifier outperformed Support Vector Machines, achieving an F1 score of 83.29%.

Huang et al. [16] investigated vocal features in depression detection. They found that patterns in vocal landmarks, like frequency and duration, effectively differentiate depressed from non-depressed speakers. Landmark pairs were particularly useful. Morales and Levitan [17] highlighted the need for a multi-signal system combining vocal and textual features for efficient depression detection. They noted that both semantic and syntactic features, along with acoustic features, are crucial for developing automated depression detection systems.

Various studies have shown the efficacy of combining multiple data types. Abhay et al. [15] focused on vocal emotion recognition, while Sharifa et al. [16] examined behavioral responses, and Yashika et al. [17] utilized EEG signals. Multimodal methods integrating audio, behavioral data, and facial expressions provide more comprehensive and accurate assessments [18]. Almars et al. demonstrated that Bi-LSTM models outperform traditional methods in detecting depression, suggesting that advanced machine learning techniques enhance detection accuracy.

Early automated depression detection attempts focused on feature isolation. Sun et al. [19] selected clinical interview questions related to mood and sleep to train a Random Forest algorithm for depression detection. Yang et al. [20] used decision trees for mental state evaluation, while Williamson et al. [21] achieved promising results with a Gaussian process model analyzing semantic text. Recent advancements include Haque et al. [22], who utilized a GRU model and BiLSTM with an attention layer for audio-text fusion, and Lin et al. [23], who applied BERT and CNN in a multimodal feature fusion framework for depression symptom estimation.

Given depression’s multifactorial nature, multimodal approaches are superior to unimodal techniques. This paper proposes an automated, early-detection model for adolescent depression, integrating EEG, ECG, and speech signals using Transformer architecture for feature extraction and multimodal data fusion.

3. METHOD

Early detection and intervention are of essence in the management of depression. This paper proposes a new approach to multimodal information fusion with the aim of high accuracy in the diagnosis of depression by considering EEG, ECG, and speech signals comprehensively. In doing so, diagnostic reliability and comprehensiveness are improved through settings extracted from a broader spectrum of information across various physiological and behavioral dimensions, hence outperforming traditional single-modality models.

In signal processing, sequential features were extracted by using a Transformer model. Self-attention mechanism forms one of the critical constituting modules of the Transformer. Hence, in any claim, it has proven helpful in handling sequence data. It cleverly pays attention to different positions in the sequence, hence more effectively capturing semantic relationships. Another critical component of this Transformer, allowing it to assess information on the dimensionality of several representational spaces, is the multi-head attention mechanism. It works by projecting inputs into different subspaces, computing attention weights corresponding to each, and then making a combination of these representations for output.

In the feature fusion part, the concat method is used to achieve the purpose of feature fusion by splicing the number of channels of the array. Compared to the additive method, this method preserves the features better and does not obscure them because of the additive.

This multimodal information fusion method is enriched by signal processing, data augmentation, and random sampling techniques that improve the generalizability of the model and reduce the challenge caused by limited data volume. This new technology opens up fresh opportunities for early depression detection and brings new hope for pushing forward research and practice in the field of mental health. The detailed block diagram of the method is shown in Figure 1.

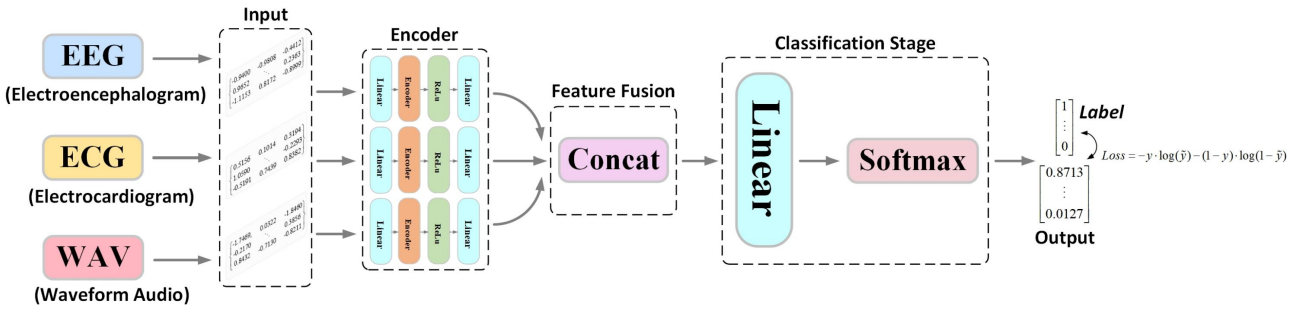

Figure 1. Structural Diagram of our method

The novelty of this study is highlighted by the fact that it creates an automated system capable of detecting incipient depression while having the ability to inspect the multimodal signal processing technology and deep learning technology in the healthcare industry. This system will fill a key gap in automatic depression detection and be a solid early warning system for clinical depression. It thereby facilitates timely intervention for patients and unequivocally marks the strong potential of deep learning technologies in healthcare, setting out a roadmap for the future evolution of depression diagnostic technologies.

4. EXPERIMENTS

We have conducted the experiment on the adolescent early-stage depression dataset from the 9th National College Student Biomedical Engineering Innovation Design Competition. Our training dataset consisted of 30 sets with signals from EEG, ECG, and voice modalities for both 22 healthy control subjects and 8 patients with depression during a reading task. There were 15 sets within the validation dataset in total, sourced from 7 healthy controls and 8 patients with depression. Experimental environment: Huawei Matebook X Pro with RTX 3090. We used the cross-entropy loss function, Adam optimizer, learning rate of 0.001, and segments of data of length 2160.

First of all, EEG, ECG, and voice signals were segmented to increase data diversity for developing a more robust model. Segmentation was not only technically demanding but also required a sense of understanding about the characteristic signal features. Each segment thus represented the physiological or speech information in a certain temporal window, enabling the model to learn across different time scales and more effectively capture subtle changes associated with

depression. These segments were later used to randomly select training data, which augmented the training dataset and further increased the model's ability to generalize to unseen data, thereby maintaining high accuracy under new scenarios.

Data augmentation is a very important technique to help models generalize better. Herein, noise is added at specified signal-to-noise ratios to signal segments; this will imitate different kinds of disturbances expected in a real environment, for example, environmental and equipment noises. This approach considerably enhances the generalizability of the model by increasing the complexity of the data and also forcing the model to learn more robust feature representations. Such improvements truly bring model robustness and stability to real-world applications. Accurate addition of noise, with its type and intensity controlled, requires very good knowledge of signal characteristics and probable sources of noise. This is very important in making sure that the model has enough preparation regarding the different noise conditions it may encounter during training and sets it up for consistent performance in real-world scenarios.

In the research, all sorts of signals, EEG, ECG, and WAV, are uniformly processed to the length of 2160 and then randomly selected for training. Only through this uniform treatment will different modalities have consistent signal handling, which is very fundamental for effective information fusion. Afterwards, the fused features are fed into a classifier for the diagnosis of depression.

To further validate the superiority of the training results with multimodal features over unimodal or bimodal feature integration, we conducted comparisons among different combinations of ECG, EEG, and WAV signals, including each paired combination and their respective solo inputs, assessing their impact on overall prediction outcomes.The accuracy of the different inputs is shown in Figure 2, and the loss value is shown in Figure 3, 4 and 5.

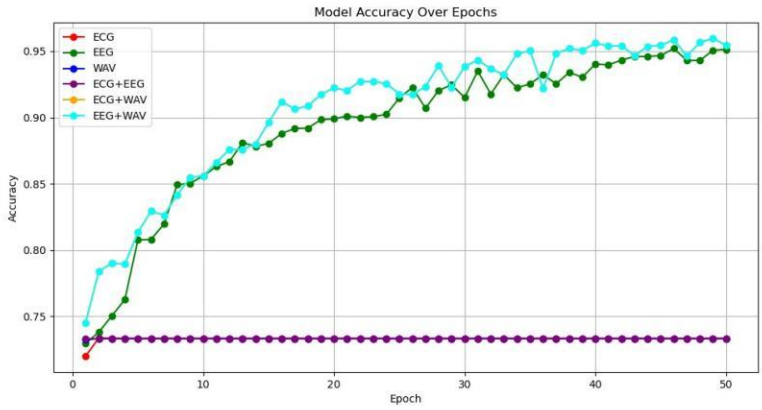

Figure 2. Accuracy of Predictions Using Single Modality or Bimodal Feature Fusion as Inputs

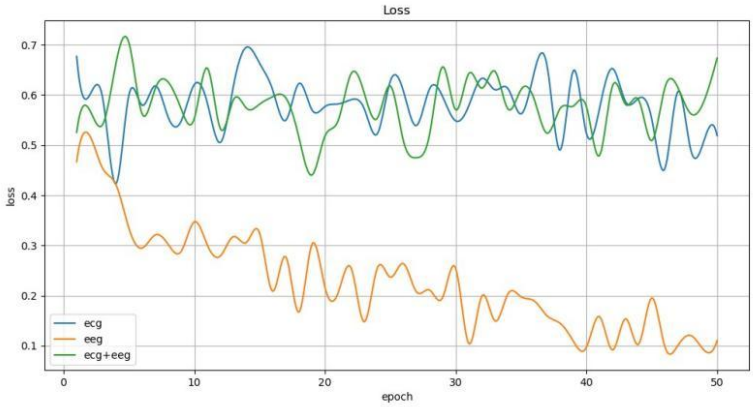

Figure 3. Loss Values for Single Modality and ECG+EEG as Inputs

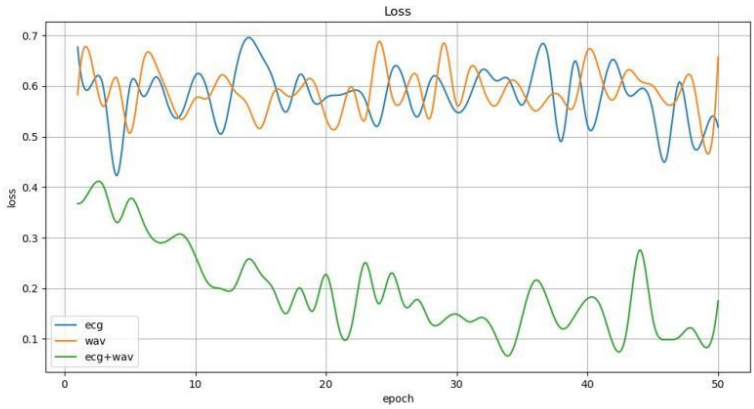

Figure 4. Loss Values for Single Modality and ECG+WAV as Inputs

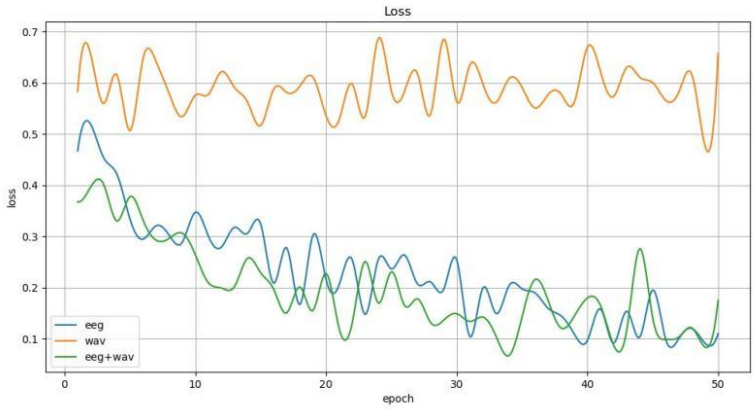

Figure 5. Loss Values for Single Modality and EEG+WAV as Inputs

Results indicate that, though each modality specifically offers some benefits, the predictive accuracy is low for a model with ECG, EEG, and WAV signals inputted separately. These unimodal predictions do not leverage potential benefits from multimodal data, of course. Combining the features of two modalities, say ECG with EEG or ECG with WAV, may likely improve predictive accuracy. These bimodal combinations have a better identification of specific physiological and psychological states compared to the performance using individual modalities.

While integrating all three modal features—ECG, EEG, and WAV—simultaneously into the model, both accuracy and stability of the predictions improve significantly. Provided with this comprehensive input, the model can become more aware of an individual's physiological and psychological states, whereby not only does prediction accuracy increase but also model robustness is enhanced. This fusion of multimodality enables the capturing of subtle feature associations for more accurate predictions.

We compared the performance of a Transformer, CNN, RNN, and Bi-LSTM for this dataset to know which models reap the best result. CNN is very renowned due to its sterling performance on image and sequential data. RNN and Bi-LSTM are much preferred for processing sequential data. At the same time, Bi-LSTM often overshadows them by better handling bidirectional dependencies. Results are shown in Figure 6 and Figure 7.

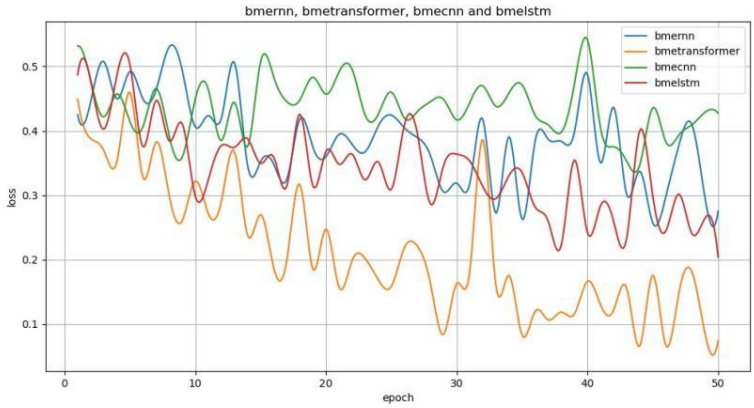

Figure 6. Loss Values for Training Using Features Extracted with Transformer, RNN, CNN, and Bi-LSTM Models

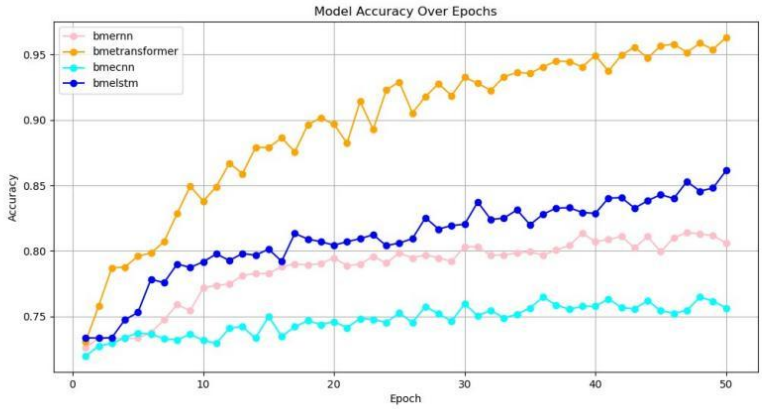

Figure 7. Accuracy of Training Predictions Using Features Extracted with Transformer, RNN, CNN, and Bi-LSTM Models

CNN models are rarely used for processing sequential data. They are majorly applied in image processing. It is not able to remember the past information regarding the inputs whilst handling long sequences due to having no memory. Nevertheless, RNN and BiLSTM models are fitted with the ability of storing the past information of the inputs within the RNN's hidden state. Not only does the BiLSTM capture previous and oncoming relationships to a signal, but it also serves as information from both past and future simultaneously, providing huge gains in the analysis of EEG, ECG, speech signals, and other sequential data. These models, however, have the undesirable issue of vanishing or exploding gradients. The Transformer model can capture dependencies regardless of the distance with a self-attention mechanism, not fearing gradient issues, and process very long sequences more effectively. Therefore, it outperforms other models on a large number of prediction tasks.

5. CONCLUSIONS

In this research, a system for the early detection of depression with the Transformer model has been addressed and developed. Our system identifies at a high degree the early stages of depression from electroencephalogram, electrocardiogram, and speech signals. Signals were then preprocessed using some meticulous techniques wherein they were randomly selected, augmented, and added with noise. Moreover, the extraction and integration of multimodal features improved the generalization performance of the model and effectively avoided the risk of overfitting, thereby ensuring stability and accuracy across different conditions.

The results from the experiments reiterated the deep impact of multimodal information fusion on improving model performance, more specifically its generalization capability. Comparative analysis of unimodal and multimodal feature performance showed that multimodal signals are much more informative about health. This methodology does not stop

at showing complex interrelationships between depression and other physiological states but forms the base for a more robust informational base for early warning systems against depression. Moreover, random selection strategies in preprocessing and noise addition significantly enhanced model robustness, substantially keeping up the high rate of recognition accuracy in the face of challenging factors like data scarcity and noise interference.

While this study has achieved some successes, a number of challenges still persist or turn out to be directions for future research. First, the data quality and quantity matter for optimal model performance. Future studies could focus on acquiring higher-quality, larger-scale multimodal data and developing more sophisticated data augmentation strategies, which might further boost model generalization. Second, improving the interpretability of the model will be highly important. Developing methods that could make a decision-making process transparent—so the professional healthcare providers could understand and trust the model's predictions—remains critical. Finally, the technological transition of this system requires practical application and clinical validation. This will be realized by applying the findings of this study in real-world situations and conducting large-scale clinical trials that can help determine its actual effectiveness and feasibility as an early depression detection system.

This work has not only introduced a new technological approach to the early diagnosis of depression but has also exposed new fields for the application of multimodal signal processing and deep learning technologies in healthcare. We are excitedly awaiting further research that can broaden and further these technologies, making large contributions to the progress of mental health technology.

REFERENCES

[1] S. Lee, J. Jeong, Y. Kwak, and S. K. Park, “Depression research: where are we now?,” Molecular Brain, vol. 3, no. 1, p. 8, 2010, doi: https://doi.org/10.1186/1756-6606-3-8.

[2] E. S. Paykel, “Basic concepts of depression,” Dialogues in Clinical Neuroscience, vol. 10, no. 3, pp. 279–289, Apr. 2022, doi: https://doi.org/10.31887/dcns.2008.10.3/espaykel.

[3] A. Thapar, S. Collishaw, D. S. Pine, and A. K. Thapar, “Depression in adolescence,” The Lancet, vol. 379, no. 9820, pp. 1056–1067, Mar. 2012, doi: https://doi.org/10.1016/s0140-6736(11)60871-4.

[4] Jia Zhiyun; Huang Xiaoqi; Wu Qizhu; Zhang Tijiang; Lui Su; Zhang Junran; Amatya Nabin; Kuang Weihong; Chan Raymond C K; Kemp Graham J; Mechelli Andrea; Gong Qiyong. High-field magnetic resonance imaging of suicidality in patients with major depressive disorder[J]. The American journal of psychiatry,2010.

[5] Juan Bueno-Notivol; Patricia Gracia-García; Beatriz Olaya; Isabel Lasheras; Raúl López-Antón; Javier Santabárbara. Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies[J]. International Journal of Clinical and Health Psychology, 2020.

[6] J. Gao et al., "Mental health problems and social media exposure during COVID-19 outbreak," PLoS ONE, vol. 15, no. 4, Art. no. e0231924, 2020.

[7] Y. Meesters, "Sensitivity to change of the Beck Depression Inventory versus the Inventory of Depressive Symptoms," J. Affect. Disord., vol. 281, pp. 338–341, 2021.

[8] Z. Wang, Z. Ma, Z. An, and F. Huang, "A Novel Diagnosis Method of Depression Based on EEG and Convolutional Neural Network," in Frontier Computing, J. C. Hung, N. Y. Yen, and J. W. Chang, Eds. Singapore: Springer, 2022, vol. 827, Lecture Notes in Electrical Engineering, pp. [page range]. doi: 10.1007/978-981-16-8052-6_10.

[9] S. Krishna and J. Anju, “Different Approaches in Depression Analysis: A Review,” 2020 International Conference on Computational Performance Evaluation (ComPE), Jul. 2020, doi: https://doi.org/10.1109/compe49325.2020.9200001.

[10] Cai H, Han J, Chen Y, et al. A pervasive approach to EEG-based depression detection[J]. Complexity, 2018, 2018(1): 5238028.

[11] M. Haque, A. Guha, and B. Schuller, "Automatic Depression Detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based Model," arXiv preprint arXiv:2202.08210, 2022. Available: https://arxiv.org/abs/2202.08210.

[12] M. Haque, A. Guha, and B. Schuller, "Automatic Depression Detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based Model," arXiv preprint arXiv:2202.08210, 2022. Available: https://arxiv.org/abs/2202.08210.

[13] Khandoker A H, Luthra V, Abouallaban Y, et al. Predicting depressed patients with suicidal ideation from ECG recordings[J]. Medical & biological engineering & computing, 2017, 55: 793-805.

[14] Ye J, Yu Y, Wang Q, et al. Multi-modal depression detection based on emotional audio and evaluation text[J]. Journal of Affective Disorders, 2021, 295: 904-913.

[15] Gong Y, Poellabauer C. Topic modeling based multi-modal depression detection[C]//Proceedings of the 7th annual workshop on Audio/Visual emotion challenge. 2017: 69-76.
[16] Lin C, Hu P, Su H, et al. Sensemood: depression detection on social media[C]//Proceedings of the 2020 international conference on multimedia retrieval. 2020: 407-411.
[17] Deshpande M, Rao V. Depression detection using emotion artificial intelligence[C]//2017 international conference on intelligent sustainable systems (iciss). IEEE, 2017: 858-862.
[18] Huang Z, Epps J, Joachim D. Investigation of speech landmark patterns for depression detection[J]. IEEE transactions on affective computing, 2019, 13(2): 666-679.
[19] Morales M R, Levitan R. Speech vs. text: A comparative analysis of features for depression detection systems[C]//2016 IEEE spoken language technology workshop (SLT). IEEE, 2016: 136-143.
[20] Orabi A H, Buddhitha P, Orabi M H, et al. Deep learning for depression detection of twitter users[C]//Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic. 2018: 88-97.
[21] Sun, Y. Zhang, J. He, L. Yu, and Q. Xu, "A random forest regression method with selected-text feature for depression assessment," in Proc. AVEC 2017, 2017, p. 61–68.
[22] L. Yang, D. Jiang, L. He, E. Pei, and M. C. Oveneke, "Decision tree based depression classification from audio video and language information," in Proc. AVEC 2016, 2016, p. 89–96.
[23] J. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, and P. Khorrami, "Detecting depression using vocal, facial and semantic communication cues," in Proc. AVEC 2016, 2016, pp. 11–18.