Application of Pre-trained Model-based Speech Analysis in Depression Detection: A Multi-task Feature Study on the MODMA Dataset

Xu Gaofeng[1], Zhou ChenYu [2]

Abstract—**Detecting depression at an early stage is critical for both public health and the well-being of patients. Even though there has been much improvement with the use of automatic depression assessment technologies based on machine learning, a number of problems still exist: demographic confounding factors; complicated feature engineering; and data privacy worries; restricted sample size; and so on. To address these challenges, this study proposes a depression detection method based on pre-trained models that utilize speech data from the MODMA dataset.**

The innovation of this study lies in the use of pre-trained models to solve problems related to small sample sizes, while adapting the model to the Chinese language contexts and enhancing its generalization capabilities. At the same time, this study systematically examines the effectiveness of verbal tasks associated with different emotions and categories in detecting depression. These findings not only improve the accuracy, practicality, and privacy protection of depression diagnosis but also offer new insights for achieving more personalized and precise mental health management. Future research will further explore the model's generalization ability across diverse datasets and aim to apply it towards developing practical applications.

Keywords—Pre-trained models, Depression detection, Speech analysis, Emotional state analysis, MODMA dataset

1. INTRODUCTION

Depression is a mental health disorder that significantly affects the global population, with its prevalence and severity demanding urgent attention. Data from the World Health Organization (WHO) indicates that around 3.8% of the global population, equating to roughly 280 million individuals, experience depression. In China, as of 2022, the prevalence of depressive disorders among adults was 3.6%, while the detection rate of depression among adolescents reached an alarming 24.6%, with severe depression accounting for 7.4%. These statistics highlight the widespread nature and urgency of depression both globally and particularly in China [1-2].

AI's integration into healthcare has recently created new avenues for addressing depression detection challenges. Scientists are now investigating various AI-aided methods, ranging from biomarker analysis to social media mining, and from video and audio processing to EEG and skin conductance measurements. Some researchers are even combining multiple data types for a more comprehensive approach [3-9]. Non-invasive techniques have become particularly appealing due to their safety, ease of use, and minimal patient disruption.

This study proposes a diagnostic approach to depression utilizing a pre-trained model. Specifically, we analyze voice data from the MODMA dataset [10] with the objective of addressing the limitations inherent in existing methodologies and enhancing both the accuracy and practicality of diagnosis.

The structure of this paper is as follows: First, the "Related Work in Depression Recognition" section provides an overview of the current state of research in automatic depression detection. Second, the "Dataset" section details the MODMA dataset used. The "Model and Experimental Setup" section introduces the model architecture used for depression recognition and provides detailed information on the experimental setup. Next, the "Experiments and Results" section analyzes the experimental results, including the impact of speech tasks with different emotional categories on detection accuracy, as well as the results of model optimization, followed by an in-depth discussion of these findings. The ‘Conclusion’ section wraps up the key findings of this study and suggests avenues for further investigation.

2. RELATED WORK ON DEPRESSION RECOGNITION

2.1. Depression Recognition Based on Single Modality

Multiple studies have demonstrated that depression symptoms can manifest through various data types, including text expression, facial expressions, and speech features. In the field of text-based analysis, Pon Karthika et al. [11] proposed an efficient method for recognizing depression from social media text data using natural language processing techniques and machine learning models. Feature extraction combined with a hybrid LSTM model was used to accomplish this. In the field of visual analysis, Y. Guo et al. [12] developed a time-extended convolutional network to extract depression-related features from video data. However, visual analysis is sensitive to environmental conditions and privacy issues despite providing rich non-verbal information. In the field of speech analysis, Zhang et al. (2021) [13] applied self-supervised learning techniques to extract both in-domain and out-of-domain audio feature embeddings, followed by the use of a BiLSTM network to predict depression levels from audio samples. The key benefit of speech analysis is that it is non-invasive and easy to collect; however, it can be influenced by factors such as environmental noise and individual differences among speakers. Additionally, physiological signal analysis has gained growing interest, with researchers like Li et al.[14] using electroencephalogram (EEG) data for depression detection, while Andy et al. [15] investigated the use of heart rate variability (HRV) for the same purpose. These methods provide direct physiological evidence but are often limited by the requirement for specialized equipment which restricts their widespread application.

2.2. Depression Recognition Based on Pre-trained Models

With advancements in self-supervised learning (SSL) technology, significant progress has been made in depression recognition methods based on pre-trained models. These methods primarily utilize large-scale unlabeled speech data to pre-train models before fine-tuning them on small-scale labeled data to address the issue of insufficient medical field data. Dumpala et al. [16] compared various SSL pre-trained models (such as wav2vec, HuBERT) in recognizing depressive symptoms, finding that these pre-trained models showed significant advantages over traditional speech features in recognizing depressive symptoms and predicting depression severity. Additionally, some studies directly used SSL pre-trained models to detect depression. These methods typically keep the weights of the pre-trained models fixed and only train a limited number of layers specific to downstream tasks, thus making efficient use of the generalized speech features learned by the pre-trained models [17] . However, most pre-trained models are primarily based on English data, and their performance in other languages (such as Chinese) needs further verification. Moreover, effectively adapting these models to specific medical tasks remains a challenge.

2.3. Multi-modal Depression Recognition

Given the constraints of single-modality data, researchers have shifted their attention to the complementary strengths of multi-modal datasets. Y. Wang et al. [18] extracted users' text, image, and behavioral data from social media to predict depression. Muzammel et al. [19] studied the impact of neural network architectures and multi-modal fusion methods on multi-modal depression diagnosis tasks. Fang et al. [20] used a multi-level attention mechanism to fuse audio, video, and text features, further improving prediction accuracy. In addition to conventional audio, visual, and text features, some innovative research, such as the team from Lanzhou University [21], attempted to analyze the kinetic and potential energy information of participants' walking collected by motion devices to predict their depression status. Although multi-modal methods significantly improve the accuracy and robustness of depression recognition, they also bring complexities in data collection and processing, as well as challenges in modality alignment and fusion.

2.4. The Impact of Speech Tasks on Depression Recognition

It is noteworthy that the nature of speech tasks may also influence the results of depression recognition. Research has shown that different types of speech tasks may lead to varying diagnostic accuracies. S.A. Almaghrabi et al. [22] found that different types of speech tasks (such as spontaneous speech, reading texts, and describing pictures) could result in different diagnostic accuracies. Additionally, both the complexity and length of the task can impact the reliability of the outcomes. Longer speech samples generally provide more information but also increase the complexity of analysis. Therefore, when designing speech-based depression recognition systems, careful consideration of the selection and design of speech tasks is necessary to optimize diagnostic accuracy and reliability.

3. DATASET

3.1. Experimental Dataset

The dataset employed in this research is the MODMA (Multi-modal Open Dataset for Mental Health Analysis), a publicly available multi-modal dataset specifically designed for mental health research, with a primary focus on depression. The audio experiments within the MODMA dataset involved a total of 52 participants, including 29

from the healthy group and 23 from the depression group, with participants ranging in age from 18 to 55 years (see Table 1).

Table 1. Demographics of male and female subjects

	Healthy controls	Depression patients
Male	20	16
Female	9	7

The experiment collected four types of speech data:

Interview Speech: Participants responded to 18 questions derived from widely recognized assessment tools, including the DSM-IV and the Hamilton Depression Rating Scale. The questions were evenly distributed into six positive, six neutral, and six negative categories.

Paragraph Reading: Participants engaged in reading a fable titled *The North Wind and the Sun*.

Word Reading:Participants read six sets of words, with each set comprising ten words. These sets were categorized into two positive, two neutral, and two negative groups.

Picture Description: Participants described four images; the first three were sourced from the Chinese Facial Affective Picture System, while the last image was obtained from the Thematic Apperception Test (TAT).

The experiment was conducted in a quiet soundproof environment where background noise levels were maintained below 60 dB. Audio recordings were made at a sampling rate of 44.1 kHz with a bit depth of 24 bits. All audio files were saved in uncompressed WAV format and manually segmented and labeled to retain only participants' speech. The dataset includes participant IDs, binary labels, PHQ-9 scores, among other relevant information.

The MODMA dataset encompasses various speech tasks along with emotional valences that provide a comprehensive reflection of participants' speech characteristics. Data collection involved both individuals diagnosed with depression as well as healthy control groups to ensure authenticity. Rigorous control over both experimental conditions and data collection processes guaranteed consistency in data quality. Furthermore, detailed metadata—including participant information and scores—are provided for thorough analysis.

3.2. Data Processing

To fully utilize the limited audio data and conduct in-depth analysis for our research objectives, we adopted a series of data processing strategies. These strategies aim to support model performance evaluation and ablation studies. First, we standardized the raw audio data. Using the FFmpeg tool [23], we unified the sampling rate of all audio files to 16 kHz and ensured the format was WAV. This step is to meet the input requirements of the speech recognition model XLSR-53 used later. Considering that the dataset is already balanced, but the audio lengths of different tasks vary greatly, we adopted targeted data segmentation and augmentation strategies:

Interview Questions: A sliding window of 5 seconds with a 2.5-second overlap was employed. This technique captures short-term emotional fluctuations while preserving contextual coherence.

Paragraph Reading: A 10-second sliding window with a 5-second overlap was implemented. This configuration allows for the retention of longer speech segments, thereby facilitating an analysis of emotional expression during reading tasks.

Word Reading:Given the relatively brief duration of each task, the entire task was treated as a single unit without further segmentation.

Picture Description: An 8-second sliding window with a 4-second overlap was utilized. This setting aims to capture potential emotional shifts during description tasks.

3.3. Data Construction

To support our research objectives, we constructed two distinct datasets:

a) Baseline Dataset: This dataset contains all the original audio segments as well as the segments divided according to the methods described above. It is used to establish the baseline performance of the model.

b) Task-Specific Dataset: The audio segments were grouped based on task type (interview, reading, description) and emotional categories. This dataset is designed to evaluate the model's performance across different tasks, supporting ablation studies. Specifically, responses to questions 1-6 were combined to form the positive group, responses to questions 7-12 were grouped as neutral, and responses to questions 13-18 were grouped as

negative. For word reading tasks, questions 20 and 23 were combined for the positive word reading group, 21 and 24 for the neutral group, and 22 and 25 for the negative group. Additionally, question 19 was used for paragraph reading, and question 29 for the thematic apperception test.

We used a stratified sampling method to divide each dataset into training, validation, and test sets in an 8:1:1 ratio. This method ensures that each subset contains a balanced representation of various task types and emotional categories. To avoid data leakage, data from the same participant was strictly limited to only one of the subsets. The relevant information regarding the preprocessing and reallocation of the baseline dataset is presented in Table 2.

Table 2. Sample distribution of the baseline dataset

	Not Depression patients	**Depression patients**
Training set	3405	2510
Validation set	427	313
Test set	427	313

4. MODEL AND EXPERIMENTAL SETUP

4.1. Speech Recognition Model XLSR-53-Chinese

In this study, we used the Chinese version of the XLSR-53 model[24] for speech recognition and feature extraction. XLSR-53, developed by Facebook AI Research, is an extension of the Wav2Vec 2.0 model that supports 53 languages, including Chinese. The core advantages of this model lie in its cross-lingual transfer learning capability and understanding of multilingual nuances.The XLSR-53-Chinese model consists of a feature encoder module, a quantization module, and a Transformer module, as depicted in Figure 1.

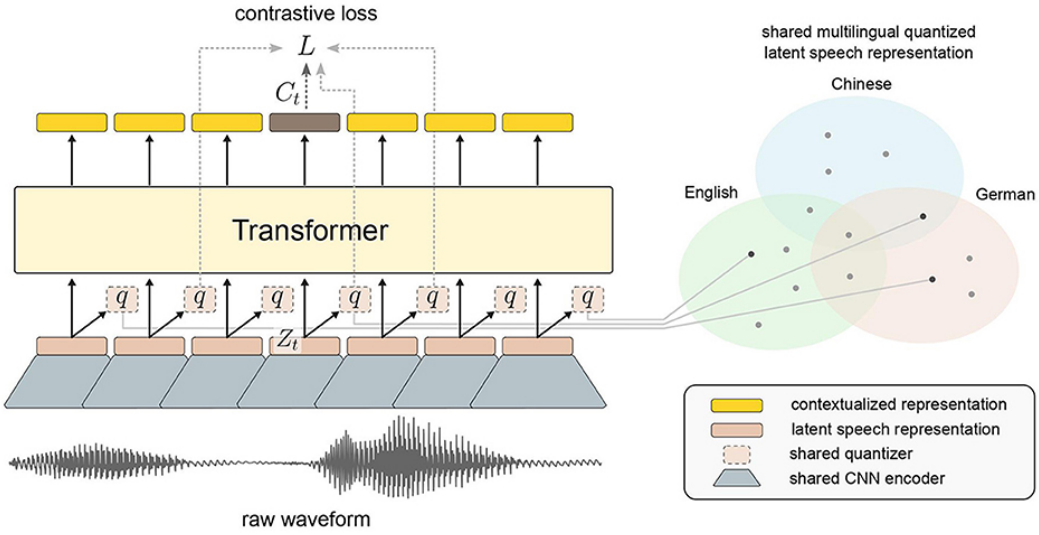


Figure 1. Framework of the wav2vec 2.0—XLSR pre-trained model.

4.2. Experimental Design and Methods

This study aims to develop a speech-based depression detection system employing a binary classification approach to identify whether subjects exhibit signs of depression. We propose a deep neural network architecture for depression detection based on the XLSR-53-Chinese pre-trained model, as illustrated in Figure 2. The architecture primarily consists of several key components:

Feature Extraction: Utilizing the pre-trained XLSR-53 model enables extraction of high-quality speech features from raw audio in a dimensionality space comprising 1024 dimensions. The XLSR-53 model is capable of automatically deriving rich representations without necessitating extensive labeled data.

Temporal Pooling: Pooling along the temporal dimension reduces feature dimensions while retaining critical information.

Fully Connected Layers: The architecture includes two fully connected layers; specifically, the first layer (self.fc1) transforms input from 1024 dimensions down to 256 dimensions, and subsequently, the second layer (self.fc2) further reduces this input from 256 dimensions down to just two dimensions—facilitating binary classification.

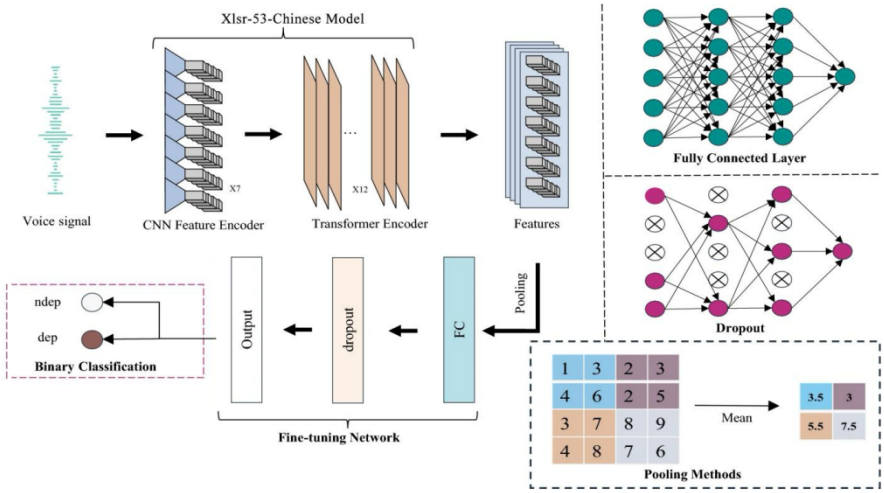


Figure 2: Architecture of voice feature detection depression system

The primary steps involved in the training process are outlined as follows: First, we initialized the fine-tuning network structure, adding task-specific layers on top of the pre-trained XLSR-53 model. The Adam optimizer was utilized, with learning rates adjusted to 1e-4, 1e-5, and 1e-6. The binary cross-entropy loss function was employed to quantify the difference between predicted values and actual labels. The calculation formula is as follows:

$$\mathrm{LCE} = -\frac{1}{N}\sum i = 1^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

The training process adopts an iterative epoch structure. At the beginning of each epoch, the model transitions into the training phase. Te training mode (model.train()) is activated, and batches of data are processed iteratively. For each batch, we perform the following operations: zero the optimizer gradients, perform forward propagation to compute model outputs, compute the loss, perform backward propagation to compute gradients, and finally update model parameters. We record and accumulate the loss of each batch to calculate the average training loss for the entire epoch.

5. EXPERIMENTS AND RESULTS

This section provides an in-depth description of our experimental procedures and analysis of the results, emphasizing the following aspects: First, we optimized model performance through hyperparameter tuning, including studying the impact of model iterations and learning rates. Next, we assessed the binary classification performance of the proposed model and benchmarked it against other established methods. Third, we conducted ablation studies to explore the contribution of different emotions and task types to depression detection. These experiments were designed to thoroughly assess the proposed model's performance on the MODMA dataset, offering key insights that could be beneficial for clinical use.

5.1. Hyperparameter Tuning

To optimize model performance, we conducted detailed hyperparameter tuning experiments, focusing mainly on the impact of model iterations (epochs) and learning rates.Using the base dataset, we set the number of training epochs to 5, 10, 20, 50, and 100, observing and recording the model performance metrics for each training epoch. Our findings indicated that as the number of epochs increased, the model's classification accuracy improved progressively. However, after the number of epochs reached 20, increasing the number of iterations no longer brought significant performance improvement. Therefore, we determined that setting the epochs to 20 was the optimal choice, achieving a good balance between performance and computational efficiency.

Next, we adjusted the learning rate of the optimizer, observing the impact of different learning rates (1 × 10^-4, 1 × 10^-5, 1 × 10^-6) on model performance. Through comparative experiments, we found that the model performed best when the learning rate was 1 × 10^-5, achieving the best balance between convergence speed and stability.

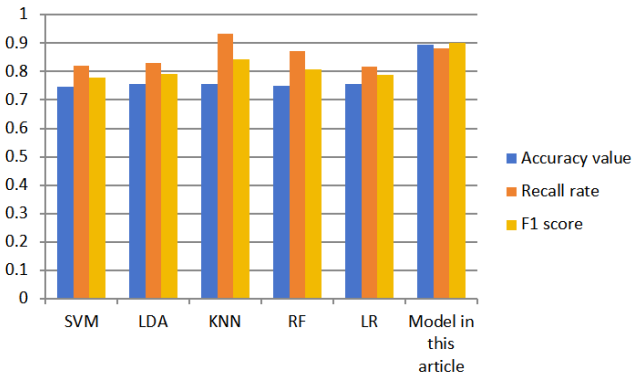
5.2. Model Performance Evaluation

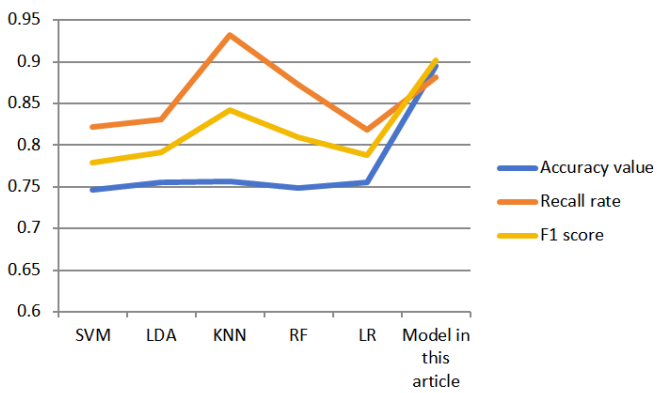
After conducting multiple experiments, Table 3 presents the classification evaluation results of six models, including accuracy, recall, and F1 score. The analysis reveals that the proposed neural network model outperforms the five classical models across all metrics. Specifically, our model achieved an accuracy of 0.895, a recall rate of 0.882, and an F1 score of 0.902, which is significantly higher than the performance of the traditional models. Two visualizations are shown below in order to further demonstrate the model performance. Figure 4a and Figure b display the bar and line charts, respectively.

Table3 shows the six models in the classification evaluation.

Table 3. Classification evaluation results of six models.

Model	**Accuracy value**	**Recall rate**	**F1 score**
SVM	0.746	0.822	0.779
LDA	0.755	0.831	0.791
KNN	0.757	0.932	0.842
RF	0.749	0.873	0.809
LR	0.756	0.818	0.788
Model in this article	0.895	0.882	0.902





(a) (b)

Figure 3. bar chart and line graph experiment Results.

5.3. Ablation Study

To gain a deeper understanding of the contribution of different task types to depression detection, we conducted detailed ablation studies. We used the base dataset with mixed emotional valences as the baseline model and then sequentially removed each component to study the effect of 12 different emotional and tasks.We found that removing the negative emotion picture description task significantly affected the accuracy, resulting in decrease in model performance. This indicates that speech tasks related to negative emotions play a crucial role in depression detection. In contrast, removing the paragraph reading task had a relatively smaller impact on the results, with accuracy only decreasing by a certain degree. This might be because the information provided by paragraph reading can be somewhat supplemented by other features.Although some features have a relatively small direct contribution to model accuracy, their importance in a clinical environment cannot be overlooked. For example, the thematic apperception test may not significantly impact model performance, but it provides rich diagnostic evidence for doctors, enhancing the model's interpretability.

6. CONCLUSION

With the utilization of speech data from the MODMA dataset as the primary source for analysis, this paper presents a novel method for depression detection employing a pre-trained model. Our experimental findings indicate that the proposed method surpasses traditional approaches, achieving an F1 score of 0.902 and an accuracy of 0.895, respectively. These results underscore the significant role of speech features in mental health diagnosis and validate the effectiveness of pre-trained models in processing speech data for depression detection.

Through model optimization and ablation studies, it became evident that various emotional categories and task types contribute differently to depression detection. Speech samples characterized by negative emotional expressions exerted the most substantial influence on detection accuracy, while neutral emotional tasks provided stable baseline features that enhanced model reliability. The integration of multiple emotional categories further improved classification performance, highlighting the importance of multidimensional emotional analysis.

However, this study has certain limitations. Although useful, the MODMA dataset was collected in a controlled environment and comprises only 189 samples, which restricts the broader applicability of our findings. The very small sample size necessitated data augmentation techniques such as speaking segmentation and merging; while these methods reduced noise levels, they may not fully capture the complexities and variations present in real-world speech data. Furthermore, given that this dataset is rooted in specific linguistic and cultural contexts, additional validation is required to assess model performance across different linguistic and cultural settings.

In summary, this research emphasize the need for a broader, more representative dataset to enhance the generalizability of the model across diverse populations. In addition, evaluating the model's performance under different emotional categories (positive, neutral, and negative) and comparing their respective contributions will help improve its practical applicability in the clinical setting of mental health diagnosis

REFERENCES

[1] W. H. Organization et al., Depressive disorder (depression), 2023.

[2] Chinese National Mental Health Development Report (2019-2020), Institute of Psychology, CAS.

[3] Wollenhaupt-Aguiar, B. et al. Differential biomarker signatures in unipolar and bipolar depression: A machine learning approach. *Aust. N. Z. J. Psychiatry* 54(4), 393–401 (2020).

[4] Li, J. et al. Intelligent depression detection with a synchronous federated optimization. *Complex Intell. Syst.* 9(1), 115–131 (2023).

[5] Casado, C. Á., Cañellas, M. L., & López, M. B. Depression recognition using remote photoplethysmography from facial videos. *IEEE Trans. Affect. Comput.* https://doi.org/10.1109/TAFFC.2023.3238641 (2023).

[6] Yang, W. et al. Attention guided learnable time-domain filterbanks for speech depression detection. *Neural Netw.* https://doi.org/10.1016/j.neunet.2023.05.041 (2023).

[7] Wang, B. et al. Depression signal correlation identification from different EEG channels based on CNN feature extraction. *Psychiatry Res. Neuroimaging* 328, 111582 (2023).

[8] Lyu, H. et al. Task-state skin potential abnormalities can distinguish major depressive disorder and bipolar depression from healthy controls. *Transl. Psychiatry* 14(1), 110 (2024).

[9] Fang, M. et al. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomed. Signal Process. Control* 82, 104561 (2023).
[10] Hanshu, Cai et al. MODMA dataset: a Multi-modal Open Dataset for Mental-disorder Analysis.
[11] Kavi Priya, S., & Pon Karthika, K. EliteVec: Feature Fusion for Depression Diagnosis Using Optimized Long Short-Term Memory Network (2023).
[12] Guo, Y. Automatic Depression Detection via Learning and Fusing Features From Visual Cues (2022).
[13] Zhang, P. Y., Wu, M. Y., Dinkel, H., & Yu, K. DEPA: Self-supervised audio embedding for depression detection. *Proceedings of the 29th ACM International Conference on Multimedia*, 135-143. Chengdu, China: ACM. DOI: 10.1145/3474085.3479236 (2021).
[14] Ksibi, A. et al. Electroencephalography-based depression detection using multiple machine learning techniques. *Diagnostics* 13(10), 1779 (2023).
[15] Schumann, A. et al. Depressive rumination and heart rate variability: A pilot study on the effect of biofeedback on rumination and its physiological concomitants. *Front. Psychiatry*, 25 August 2022, Sec. Public Mental Health.
[16] Dumpala, S. H. Self-Supervised Embeddings for Detecting Individual Symptoms of Depression (2024).
[17] Huang, X. et al. Depression recognition using voice-based pre-training model. *Sci. Rep.* 14(1), 12734. DOI: 10.1038/s41598-024-63556-0 (2024).
[18] Wang, Y., Wang, Z., Li, C., Zhang, Y., & Wang, H. Online social network individual depression detection using a multitask heterogeneous modality fusion approach. *Inf. Sci.* 609, 727–749 (2022).
[19] Muzammel, M. End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis (2021).
[20] Fang, M., Peng, S. Y., Liang, Y. J., Hung, C. C., & Liu, S. H. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomed. Signal Process. Control* 82, 104561 (2023). DOI: 10.1016/j.bspc.2022.104561.
[21] Sun, S. T., Chen, H. Y., Shao, X. X., Liu, L. L., Li, X. W., & Hu, B. EEG-based depression recognition by combining functional brain network and traditional biomarkers. *Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine*, 2074-2081. Seoul, Korea (South): IEEE (2020).
[22] Clark, S. R. et al. Bio-acoustic features of depression: A review. *Biomed. Signal Process. Control* 85, 105020 (2023).
[23] FFmpeg. https://ffmpeg.org/.
[24] Hugging Face. Wav2Vec2-large-XLSR-53-Chinese-zh-cn