Hindawi
International Journal of Intelligent Systems
Volume 2024, Article ID 1564574, 13 pages
https://doi.org/10.1155/2024/1564574

Research Article

FPT-Former: A Flexible Parallel Transformer of Recognizing Depression by Using Audiovisual Expert-Knowledge-Based Multimodal Measures

Yifu Li,[1,2] **Xueping Yang**,[3] **Meng Zhao**,[1,2] **Zihao Wang**,[1,2] **Yudong Yao**,[4] **Wei Qian**,[1] **and Shouliang Qi**[1,2]

[1]*College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China*
[2]*Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang, China*
[3]*Department of Psychology, The People's Hospital of Liaoning Province, Shenyang, China*
[4]*Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, USA*

Correspondence should be addressed to Shouliang Qi; qisl@bmie.neu.edu.cn

Received 3 September 2023; Revised 20 December 2023; Accepted 15 January 2024; Published 29 January 2024

Academic Editor: Said El Kafhali

Background and Objective. Currently, depression is a widespread global issue that imposes a significant burden and disability on individuals, families, and society. Deep learning (DL) has emerged as a valuable approach for automatically detecting depression by extracting cues from audiovisual data and making a diagnosis. PHQ-8 is considered a validated diagnostic tool for depressive disorders in clinical studies, and the objective of this experiment is to improve the accuracy of PHQ-8 prediction. Furthermore, this paper aims to demonstrate the effectiveness of expert knowledge in depression diagnosis and discuss a novel multimodal network architecture. *Methods*. This research paper focuses on multimodal depression analysis, proposing a flexible parallel transformer (FPT) model capable of extracting data from three distinct modalities (i.e., one video and two audio descriptors). The FPT-Former model incorporates three paths, each using expert-knowledge-based descriptors from one modality as inputs. These descriptors are represented into 32 features by the encoder part of a transformer module, and these features are fused to realize the final regression of PHQ-8 score. The extended distress analysis interview corpus (E-DAIC) is an expansion of WOZ-DAIC which comprises semiclinical interviews intended to assist in the diagnosis of psychological distress conditions. It encompasses a sample size of 275 participants, and in this study, it was utilized to test the model in a way of 10-fold cross-validation. *Results*. The FPT presented herein achieved comparable performance to the state-of-the-art works, with a root mean square error (RMSE) of 4.80 and a mean absolute error (MAE) of 4.58. The ablation experiments demonstrate that the three-modality-fused model outperforms other two-modality-fused and single-modality models. While using a PHQ-8 score threshold of 10, the accuracy of the depression classification is 0.79. *Conclusions*. Leveraging the strength of expert-knowledge-based multimodal measures and parallel transformer structure, the FPT model exhibits promising performance in depression detection. This model improved the accuracy of depression diagnosis through audio and video, and it also proved the effectiveness of using expert-knowledge in the diagnosis of depression. The traits of flexible structure, high predictive efficiency, and secure privacy protection make our model a promotable intelligent system in mental healthcare.

1. Introduction

The recognition that mental disorders are significant contributors to the burden of disease is growing [1]. Currently, depression stands as the most prevalent mental illness, characterized predominantly by persistent and long-term feelings of low mood [2], making it a significant form of mental illness in modern times. As per the World Health Organization (WHO), by 2030, depression is projected to become the most prevalent mental disorder [3]. In extreme cases, depression can result in suicide [4]. At present, there is no distinct and effective clinical definition for depression,

resulting in a diagnosis process that can be both subjective and lengthy. The integration of artificial intelligence and mathematical modeling methods is increasingly being employed in mental health research in attempt to address this issue. These techniques can be beneficial to the field of depression detection, given their ability to appreciate the significance of acquiring detailed data to distinguish various depression disorders [5].

A multitude of automatic depression estimation (ADE) systems have been developed [6]. Many audiovisual features can also be used to diagnose depression [7, 8]. Zhou et al. [9] put forward a unique deep architecture named Depress Net, designed to learn representations from images for depression recognition. He et al. [10] proposed a network that integrates 2D-CNN networks and an attention mechanism for depression recognition. While the majority of earlier research concentrated on single-modal data, recent studies have demonstrated that multimodal data can provide superior predictive performance compared to single-modal data [11]. Yang et al. [12] introduced a multimodal fusion framework that integrates deep convolutional neural network (DCNN) and deep neural network (DNN) models. This model uses audio, video, and text streams as inputs and is aimed at detecting depression. However, how to better mine serialized information, how to better utilize multimodal information, and which features can improve the accuracy of diagnosis more effectively are still topics that need further research.

Using knowledge-based descriptors as inputs can be an alternative strategy while using original audio and video as inputs faces the challenge of personal privacy disclosure, and the sheer volume of raw video can also slow down predictive efficiency. Facial expressions and the acoustic characteristics of speech are the two main categories of knowledge-based measures. Facial expressions serve as a powerful means of conveying emotions to others. Psychologists have meticulously modeled these expressions, culminating in a reference guide known as the facial action coding system (FACS) [13] which will be used in this experiment. This system catalogs the combinations of facial muscles involved in each expression and can be utilized as a tool to discern an individual's emotional state through their facial expression. The acoustic characteristics of speech have also been recognized as potential indicators of depression [14]. In this study, a novel set of acoustic features known as the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) has been used. This set was recently devised for use in various areas of speech analysis [15]. The Mel-Frequency Cepstral Coefficient (MFCC) method is also one of the foremost techniques used for the extraction of speech features [16], and this feature will also be used in the model proposed in this article.

Numerous open-source datasets about depression are available, offering a valuable resource for researchers delving into the realm of mental health. Shen et al. [17] presented the Emotional Audio-Textual Depression Corpus (EATD-Corpus), an assemblage encompassing audio recordings and transcribed responses gathered from both depressed and nondepressed individuals. However, it is important to note that this corpus exclusively contains audio and textual data. Cai et al. [18] introduce a multimodal open dataset designed for the analysis of mental disorders. This dataset incorporates EEG and audio data sourced from clinically depressed patients, as well as corresponding data from unaffected control subjects but its sample size is slightly above 50. The dataset used in this paper is the Extended Distress Analysis Interview Corpus (E-DAIC) [19], an enhanced version of WOZ-DAIC, which comprises semi-clinical interviews intended for assisting in the diagnosis of psychological distress conditions. These resources stem from the Audio/Visual Emotion Challenge and Workshop (AVEC 2019), an event dedicated to the comparison of multimedia and machine learning techniques in the realm of health and emotion analysis [19].

Drawing inspiration from the potent learning capacity of transformers, a sequence transduction model wholly based on attention mechanisms [20], we proposed this FPT-Former framework. This model is composed of multiple parallel encoders for each modality, which create low-dimensional global feature vectors encapsulating compact information. By combining with expert knowledge, this model enhances the prediction accuracy of the PHQ-8 scores [21]. FPT-Former is specifically tailored to process diverse data types, facial expressions, audio-MFCC, and audio-eGeMAPs, for accurate depression severity estimation. Our model surpasses existing methodologies by incorporating a confluence of components, including parallel transformer encoders for each modality and a fusion layer for effective information convergence. Our FPT-Former achieved comparable performance to the state-of-the-art works, with a root mean squared error (RMSE) of 4.80 and a mean absolute error (MAE) of 4.58.

The key contributions of this paper can be encapsulated as follows: (1) a flexible parallel transformer model has been proposed for depression recognition; (2) the fusion of audiovisual expert-knowledge-based multimodal metrics increases prediction accuracy; (3) the paralleled structure adapts different numbers of measures in diverse modalities; and (4) the utilization of low-dimensional video features in the proposed transformer model increases prediction efficiency and avoids personal privacy leakage.

The structure of the paper is organized as follows: Section 2 introduces the related work of automatic depression diagnosis. Section 3 introduces the expert-knowledge-based features that serve as inputs to the FPT-Former. Section 4 illustrates the framework of the proposed FPT-Former. Section 5 presents and analyzes the experimental results. Section 6 concludes the paper and discusses potential future work.

2. Related Work

In the area of depression recognition, many researchers have made progress. Du et al. [22] introduced the machine speech chain model for depression recognition (MSCDR), highlighting the significance of vocal tract changes as important markers for depression. Yang et al. [23] addressed the challenge of speech depression detection by proposing the

DALF framework, which employs attention-guided learnable time-domain filterbanks. By learning acoustic features and spectral attention, DALF outperformed state-of-the-art methods from audio signals. In a similar vein, Niu et al's "Depressioner" model [24] turned its attention to facial dynamics, identifying facial changes as potential biomarkers for depression levels. In 2022, Kakuba et al. [25] introduced the concurrent spatial-temporal and grammatical (CoSTGA) model, which is a deep learning-based approach. This model is designed to simultaneously acquire spatial, temporal, and semantic representations within the local feature learning block (LFLB). These representations are then combined into a latent vector, which serves as the input for the global feature learning block (GFLB), and they also presented an attention-based multilearning model (ABMD) [26] that leverages residual dilated causal convolution (RDCC) blocks and dilated convolution (DC) layers featuring multihead attention. The ABMD model delivers comparable performance while efficiently capturing global contextualized long-term dependencies among features in a parallel manner which can be used in speech emotion recognition.

Moreover, researchers have been proactive in embracing multimodal approaches to elevate the accuracy of depression recognition. Li et al's multimodal hierarchical attention (MHA) model [27], designed for social media settings, improves recognition performance by integrating various data types. This model employs attention mechanisms and combines multiple data sources which highlight the significance of taking a holistic approach. In tandem, privacy concerns were addressed by Pan et al's AVA-DepressNet [28]. This model pays attention to facial privacy preservation while concurrently boosting audiovisual feature enhancement, addressing the ethical issues of technology application in sensitive domains such as mental health. Zou et al. [29] developed a Chinese Multimodal Depression Corpus (CMDC) by conducting semistructural interviews with depression patients. Through feature analysis and benchmark evaluations, they established the effectiveness of their multimodal fusion approach, showcasing its potential for automatic depression screening. Zhang et al. [30] introduced a two-stream deep network within a depression detection framework which achieved state-of-the-art performance on AVEC2014 datasets [31].

Extending the scope, Zhao et al's work [32] pushed the boundaries of depression detection through the analysis of image-based data. They introduced frequency attention, tapping into the distinctive traits of depression images to uncover significant patterns of depression patients.

Besides, the temporal dimension emerges as a recurring motif. He et al's DepNet [33] used deep learning to extract spatial-temporal patterns from video-based facial sequences. By scrutinizing patterns over time, this approach offers a deeper understanding of the dynamic nature of depression manifestations.

Lastly, the treatment of mental illnesses is also a field of interest for deep learning. A. Singh et al. proposed a cost-effective, socially designed robot named "Tinku" for teaching and assisting children with autism spectrum disorder [34]. In addition, the continuous introduction of new deep learning models, as well as improvements and adjustments to these models, has played a promotive role in the research of this field [35–37].

3. Feature Descriptors

The data utilized in this study are derived from the Extended Distress Analysis Interview Corpus (E-DAIC) [19], a broader version of WOZ-DAIC, which comprises semi-clinical interviews intended for assisting in the diagnosis of psychological distress conditions. The research for this paper involves 275 subjects, with each participant contributing three unique sets of data, specifically FACS, eGeMAPS, and MFCC. Every subject is assigned a Patient Health Questionnaire (PHQ-8) score. All the initial descriptors used in this paper are grounded in expert-based knowledge.

3.1. Video Descriptors. The video descriptors in this experiment were extracted by a facial behavior analysis toolkit called OpenFace which can accurately detect head pose, recognize facial action units, and estimate eye gaze [38]. Each subject's interview video undergoes processing with OpenFace. After this processing, every frame of the video contains 49 distinct feature values.

The first to sixth eigenvalues constitute the subject's head pose. The orientation of the head in relation to the camera can be represented through rotation and shift. Figure 1 illustrates the description of a head rotation transformation. The spatial coordinates corresponding to the head's position will also be provided.

Features from the 7th to the 14th represent the subjects' eye gaze estimation. The gaze direction vector of each eye is represented by three numbers and the average of the horizontal and vertical radians of the gazing directions of the two eyes will also provide two features.

The Facial Action Coding System (FACS) [13] is a taxonomy of human facial movements defined by their manifestation on the face. It encodes the movements of individual facial muscles, discerning subtle instantaneous changes in facial appearance. With FACS, nearly any anatomically possible facial expression can be coded, breaking it down into specific action units (AUs). It serves as a widely accepted standard for objectively describing facial expressions [39]. Eighteen action units were considered in this study, all of which are more typically associated with the expression of negative emotions, and these AUs are described in Table 1.

Each AU is denoted by two values: its presence and its intensity (excluding AU28, for which only its presence is determined.). The presence refers to whether the AU is visibly apparent on the face and the intensity refers to the strength or force of the AU, rated on a 5-point scale ranging from minimal to maximal.

3.2. Audio Descriptors: MFCC. The first set of audio expert-knowledge-based measures is the Mel-Frequency Cepstral Coefficients (MFCC), and it encompasses 39 features.

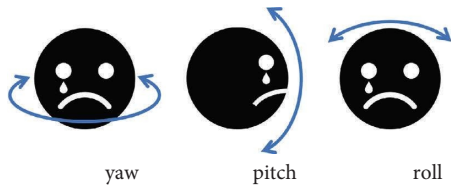

Figure 1: Relationship between head posture angle change and head motion.

Table 1: List of AUs in OpenFace.

AU	Full name
AU 1	INNER BROW RAISER
AU 2	OUTER BROW RAISER
AU 4	BROW LOWERER
AU 5	UPPER LID RAISER
AU 6	CHEEK RAISER
AU 7	LID TIGHTENER
AU 9	NOSE WRINKLER
AU 10	UPPER LIP RAISER
AU 12	LIP CORNER PULLER
AU 14	DIMPLER
AU 15	LIP CORNER DEPRESSOR
AU 17	CHIN RAISER
AU 20	LIP STRETCHED
AU 23	LIP TIGHTENER
AU 25	LIPS PART
AU 26	JAW DROP
AU 28	LIP SUCK
AU 45	BLINK

Research in psychophysics has revealed that the human perception of frequency in speech signals does not adhere to a linear scale. Therefore, for each tone with a factual frequency (f) measured in hertz (Hz), a subjective pitch is assessed on a scale known as the "Mel" scale which is calculated as follows [40]:

$$f_{\text{mel}} = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{1}$$

In this context, f_{mel} corresponds to the subjective pitch in the "Mel" scale that is associated with a specific frequency measured in hertz (Hz). This understanding forms the foundation for the definition of the Mel-Frequency Cepstral Coefficients (MFCCs), a fundamental set of acoustic features used in speech and speaker recognition applications [41].

In the data employed for this experiment, 13 coefficients are preserved following the Discrete Cosine Transform (DCT), and the first and second derivatives of these 13 coefficients are also computed. In summary, the Mel-Frequency Cepstral Coefficient (MFCC) expert-knowledge-based dataset comprises a total of 39 features.

3.3. Audio Descriptors: eGeMAPS. Valuable information encapsulating emotional indicators can be extracted from audio signals, which can contribute significantly to the diagnosis of depression. Researchers are investigating various dimensions such as the identification of emotional states, the conveyance of emotional signals through voice, the impact of emotions in language, and the automatic detection of speaker emotions for enhancing depression prediction efficacy.

eGeMAPS is an expanded version of GeMAPS [15], augmenting its set of 18 low-level descriptors (LLDs) with several new features. These additions include five spectral features, such as the first four Mel-Frequency Cepstral Coefficients (MFCC1-4), and the spectral difference between consecutive frames, alongside two frequency-dependent attributes: the bandwidth of the second and third formants. eGeMAPS contains 88 features in total which extract acoustic parameters from speech to understand vocal emotional expressions.

This paper uses 23 of these 88 features that are most relevant to the diagnosis of depression, and every feature undergoes a smoothing process utilizing a window size of three frames. Detailed descriptions of these 23 features can be seen in Table 2.

4. Framework of Flexible Parallel Transformer

The network introduced in this study is called the FPT-Former, which is a flexible, parallel transformer network specifically built to process multimodal data, facial expressions, audio-MFCC, and audio-eGeMAPs characteristics, for depression identification. The FPT-Former architecture is constituted by a confluence of components. These components encapsulate an input layer, a cohort of parallel encoders for each modality, and a fusion layer to converge the learned information across the modalities. The total number of training parameters for the entire model is 234,463, and a ReduceLROnPlateau scheduler is used to adjust the learning rate based on the validation loss, enhancing the model's ability to converge to the optimal solution. This section provides a detailed exposition of the network architecture, and the data structure in each phase undergoes various transformations throughout the pipeline of our framework, FPT-Former, as illustrated in Figure 2.

4.1. Data Preprocessing and Input. Our dataset contains multimodal data including facial expression, audio-MFCC, and audio-eGeMAPs measures. The facial expression features, derived from the FACS, consist of 49 dimensions, and audio-eGeMAPS measures comprise 39 dimensions, while audio-MFCC measures are characterized by 23 dimensions. To preserve the temporal information across the sequence of video or audio frames, an additional feature value indicating the frame serial number is appended to each modality, resulting in the dimensions of 50, 40, and 24, respectively (the feature number of facial expression, audio-MFCC, and audio-eGeMAPs measures). Each of the three modalities takes a frame every 0.1s, and each subject takes 4,146 frames.

4.2. Input Layer. The journey of data through the network commences at the input layer. Herein, the raw multimodal data are introduced into the system frame-by-frame. This modality-specific data include facial measures of dimension

Table 2: Selected eGeMAPS features for depression diagnosis.

Feature name	Feature declaration
Loudness	The overall volume or sound intensity of a sound signal
Alpha ratio	The energy ratio of the sound signal spectrum's low and high-frequency parts
Hammarberg index	The change pattern of the fundamental frequency in the sound signal
Slope 0–500	The frequency spectrum's rate of alteration is assessed in the range from 0 Hz to 500 Hz
Slope 500–1500	The frequency spectrum's rate of alteration is assessed in the range from 500 Hz to 1500 Hz
Spectral flux	The amount of flow or variation in the spectrum of a sound signal
mfcc1	The first Mel-Frequency Cepstral Coefficients
mfcc2	The second Mel-Frequency Cepstral Coefficients
mfcc3	The third Mel-Frequency Cepstral Coefficients
mfcc4	The fourth Mel-Frequency Cepstral Coefficients
F0semitoneFrom27.5 Hz	The semitone difference between the fundamental frequency of the sound signal and 27.5 Hz
Jitter local	The local jitter of the sound signal
Shimmer local dB	The local trill of a sound signal
HNRdBACF	Harmonic to noise ratio (HNR) of a sound signal
LogRelF0-H1-H2	The logarithmic difference between the fundamental frequency in the sound signal and the corresponding first (H1) and second (H2) harmonics
logRelF0-H1-A3	The logarithmic difference between the fundamental frequency in the sound signal and the corresponding first harmonic (H1) and third formant (A3)
F1frequency	The frequency of the first formant (F1) in the sound signal
F1bandwidth	The bandwidth of the first formant (F1) in the sound signal
F1amplitudeLogRelF0	The logarithmic difference between the amplitude of the first formant (F1) in the sound signal and the fundamental frequency
F2frequency	The frequency of the second formant (F2) in the sound signal
F2amplitudeLogRelF0	The logarithmic difference between the amplitude of the second formant (F2) in the sound signal and the fundamental frequency
F3frequency	The frequency of the third formant (F3) in the sound signal
F3amplitudeLogRelF0	The logarithmic difference between the amplitude of the third formant (F3) in the sound signal and the fundamental frequency

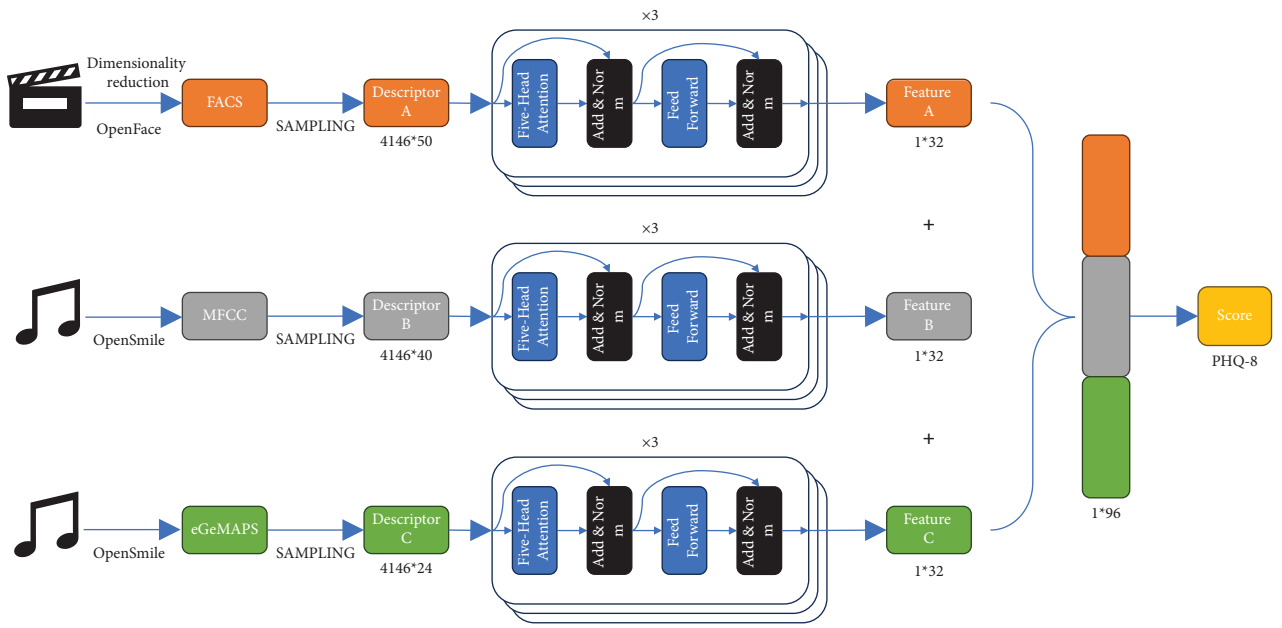

Figure 2: Framework of flexible parallel transformer.

(4146, 50), audio-MFCC measures of dimension (4146, 40), and audio-eGeMAPs measures of dimension (4146, 24). The network takes three input streams with dimensions 50, 40, and 24, respectively, corresponding to the different modalities of the dataset. This approach is predicated on the understanding that capturing the temporal dynamics

inherent in the frame sequence is paramount to the effectiveness of the model.

4.3. Encoder Stage. Following that the input layer is the encoder stage, this stage is characterized by a trident of parallel encoders, each designed to cater to the specific modalities: facial expression, audio-MFCC, and audio-eGeMAPs measures.

4.3.1. Transformer and Self-Attention. The Transformer architecture, proposed by Vaswani et al. [20], has emerged as a groundbreaking paradigm in sequential data modeling, thanks to its innovative self-attention mechanism. This mechanism allows the model to weigh the significance of different positions within a sequence while processing each element, making it well-suited for capturing long-range dependencies and relationships.

The essence of the self-attention mechanism lies in the QKV (Query-Key-Value) mechanism, which can be mathematically expressed as Figure 3.

Given an input sequence of vectors $X = (x_1, x_2, \ldots, x_i)$, the self-attention mechanism calculates the weighted sum of value vectors based on their relevance to a query vector:

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where Q represents the query matrix, K represents the key matrix, and V represents the value matrix. d_k is the dimension of the key vectors.

The scaled dot-product operation inside the softmax function computes the compatibility between each query and key pair. The result is a set of attention scores that determine how much each value contributes to the final output.

In the context of our model, the self-attention mechanism enables the encoder to focus on relevant features within a sequence. This is particularly beneficial when processing multimodal data, as it helps capture meaningful interactions between different elements.

4.3.2. Encoder Components. In this architecture, data from three different modalities are individually channeled into three separate encoder pathways. Each pathway follows a multistep process.

Initially, the data are subjected to a multihead self-attention mechanism, where the model utilizes five attention heads to capture intricate relationships and dependencies within each modality. Then, a layer normalization step is applied after the multihead self-attention process. The Transformers are tailored to the specific requirements of each data modality, with different numbers of heads and layers. The model uses 5, 4, and 4 self-attention heads in its three transformer modules, respectively. Following layer normalization, the data passes through position-wise feed-forward networks. To mitigate the risk of gradient vanishing, residual connections are employed and these connections allow the output of each layer to be combined with its input. Three identical encoder layers are stacked one upon the other. The number of encoder layers and self-attention heads is determined after multiple attempts under our computational environments. Each encoder layer encompasses all the aforementioned components, and this stacking increases the model's representational capacity. At the end of this process, each pathway yields an output vector with dimensions [1, 32].

4.4. Fusion Layer. After the three encoders produce their respective outputs, three [1, 32] feature vectors are obtained. These vectors are then combined in a fusion layer, resulting in a single comprehensive feature vector of dimensions (1, 96). This aggregated vector encompasses all the essential information from the three modalities. Subsequently, the feature vector is passed through a fully connected layer, culminating in the final PHQ-8 score which can be used in the prediction of depression severity.

In summary, the FPT-Former utilizes the richness of information intrinsic to the different modalities, enabling their synergistic utilization to augment the prediction accuracy of depression severity. The next section will shed light on the effectiveness of our model, substantiated by empirical results from our experiments.

5. Experiments and Analysis

In this section, we will analyze the experimental results, assess our proposed FPT-Former model, and conduct a comparison with existing state-of-the-art techniques. Furthermore, through ablation studies, we will substantiate the effectiveness of the FPT-Former model in estimating depression severity by expert-knowledge-based multimodal measures.

5.1. Dataset Split and Model Training. This study makes use of the E-DAIC dataset which comprises data from 275 participants. Each participant's data represents a sample, which includes visual features from OpenFace 2.1.0, eGeMAPS features, and MFCC features extracted using OpenSMILE [42]. For each sample, the maximum frames considered are 12,438 for visual data and 41,460 frames for both MFCC data and eGeMAPS features.

To ensure the consistency of the input across all samples, we restrict the data for each modality. For visual features, we select every third frame, resulting in 4146 frames per sample. For audio features, every tenth frame is chosen, resulting in 4146 frames per sample for each of these modalities as well.

We employ a ten-fold cross-validation scheme for our model training and evaluation, thus splitting our dataset into ten partitions. For each fold, nine partitions are used for training, and one partition is left out for testing.

The FPT-Former is trained using the Adam optimizer with an initial learning rate of 0.01 and a cosine annealing schedule for learning rate decay. The model is evaluated using two metrics: RMSE and MAE, and both of them have been calculated on the validation set for each fold.

RMSE and MAE are defined as follows:

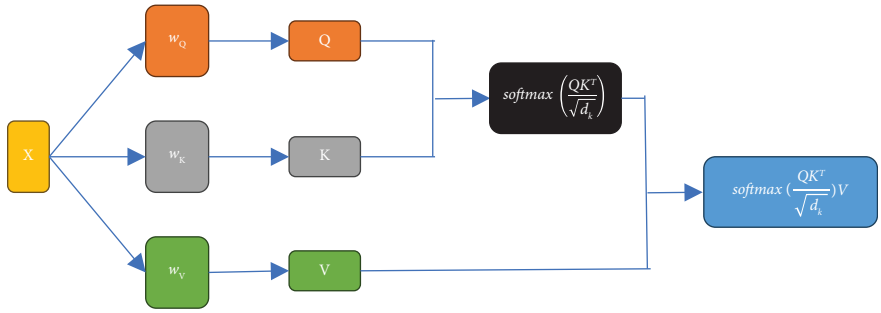

Figure 3: Self-attention mechanism.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum (r_i - r_i')^2}, \qquad \text{MAE} = \frac{1}{N}\sum |r_i - r_i'|, \tag{3}$$

where N is the total number of observations, r_i is the prediction from the model, and r_i' is the actual observed value.

Each fold is repeated and the reported results are averaged over all folds. In this manner, we ensure a robust estimation of our model's performance.

5.2. Depression Recognition Results. To establish the effectiveness of our proposed FPT-Former model in multimodal depression severity estimation, we compared it with several existing state-of-the-art methods. The comparative evaluation focused on the primary performance metrics: RMSE and MAE.

Table 3 outlines the performance of our method against others. Al Hanai et al. [43] employed audio and text features in an LSTM neural network, achieving an RMSE of 6.50 and MAE of 5.13, and Zhang et al. [44] introduced an autoencoder model with BiGRU for speech-based depression severity prediction, resulting in an RMSE of 5.68 and MAE of 4.64. Yang et al. [45] integrated speech, text, and face data with DCGAN for feature augmentation, yielding an RMSE of 5.52 and MAE of 4.63, and Han et al. [46] proposed a spatial-temporal feature network for speech-based depression detection, achieving an RMSE of 6.29 and MAE of 5.38 while Fang et al. [47] presented a multimodal fusion model with multilevel attention mechanism for depression detection, with an RMSE of 5.17. Our FPT-Former presented herein achieved comparable performance to the state-of-the-art works, with an RMSE of 4.80 and an MAE of 4.58. To ensure comparability, all the methods listed in Table 3 used E-DAIC as a dataset.

To provide a better understanding of the agreement between our FPT-Former model's predictions and the actual depression severity scores, we conducted a Bland–Altman analysis. The Bland–Altman plot (the left part of Figure 4) depicts the difference between the predicted depression severity scores and the actual scores on the y-axis, against the average of the two scores on the x-axis. The regression analysis (the right part of Figure 4) also allows us to observe whether the model exhibits consistent deviations across different levels of depression severity.

5.3. Ablation Experiment. To better understand the contribution of each modality to our model's performance, we conducted ablation experiments. These experiments systematically removed one or two modalities from the multimodal model and observed the effect on performance.

5.3.1. Single Modality Ablation. In the single modality ablation experiments, we individually removed each modality such as FACS, MFCC, and eGeMAPS from our FPT-Former model and observed the change in model performance. Table 4 presents the results of the ablation experiments, showing the RMSE and MAE values when each modality was removed.

From the results presented in Table 4, it can be observed that each modality plays a vital role in the performance of the FPT-Former model. Removing any one of the modalities leads to an increase in RMSE and MAE, indicating a decline in prediction accuracy. This underlines the importance of multimodal data and the synergy between these modalities in making accurate predictions. The extent of performance degradation varies with the removal of different modalities, suggesting that each modality contributes differently to the overall model's performance.

5.3.2. Double Modality Ablation. Next, we examined the interplay between different modalities by conducting double modality ablation experiments. Here, we removed two modalities at a time and evaluated the performance of the model with only the one remaining modality. The results are shown in Table 5.

These findings reinforce the notion that each expert-based-knowledge carries unique and valuable information for the task of depression severity estimation. Relying on a single modality can cause the loss of essential information. This underlines the significance of an expert-based-knowledge in developing robust predictive models for depression recognition. To visually represent these findings, a bar plot was generated (Figure 5) to compare the RMSE and MAE values for different ablation scenarios. The plot illustrates the impact of removing each modality on the model's predictive accuracy.

The dataset we utilized originates from AVEC 2017: Real-life Depression and Affect Recognition Workshop and Challenge [19], and this challenge provided a baseline for comparison. We compared the results of our double

Table 3: The performance of depression recognition on E-DAIC databases.

Study	Model name	Modality	RMSE	MAE
Al Hanai et al. 2018 [43]	Long-short term memory (LSTM) neural network	Audio and text features	6.50	5.13
Zhang et al. 2020 [44]	An autoencoder model based on a bidirectional gated recurrent unit (BiGRU)	Speech signals	5.68	4.64
Yang et al. 2020 [45]	Deep convolutional generative adversarial network (DCGAN)	Speech, text, and face data	5.52	4.63
Han et al. 2023 [46]	Spatial-temporal feature network (STFN)	Speech data	6.29	5.38
Fang et al. 2023 [47]	A multimodal fusion model with a multilevel attention mechanism (MFM-Att)	Audiovisual and text data	5.17	—
Ours	A flexible parallel transformer model (FPT-Former)	Audiovisual expert-knowledge-based measures	**4.80**	**4.58**

The bold font indicates the lowest value among the compared studies.

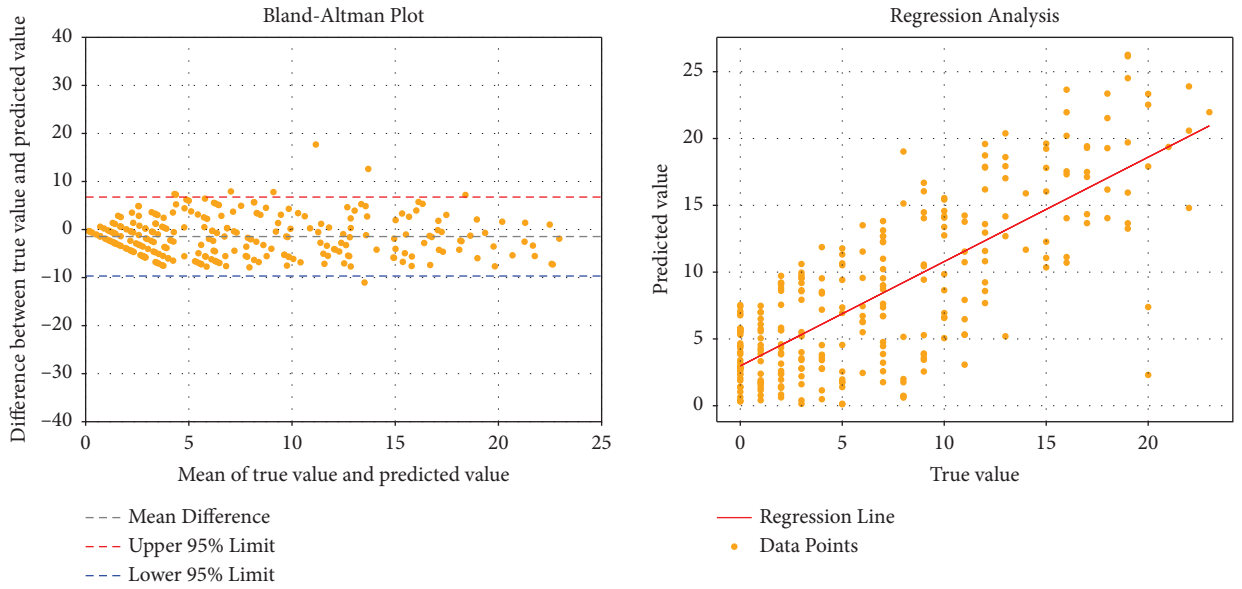

Figure 4: The Bland–Altman plot and regression analysis for FPT-Former depression severity predictions.

Table 4: The performance of FPT-Former when each modality is removed.

Model variant	RMSE	MAE
Without FACS	6.02	5.86
Without MFCC	5.67	5.51
Without eGeMAPS	5.88	5.73

Table 5: The performance of FPT-Former when only one modality is used.

Model variant	RMSE	MAE
FACS only	6.11	5.91
MFCC only	7.02	6.84
eGeMAPS only	6.60	6.21

modality ablation experiment with the baseline (the challenge provided the RMSE only) from the challenge, and the results (Table 6) showed that our model demonstrated better performance in predictions.

The results reveal that the removal of FACS has the most significant impact, leading to the highest increase in both RMSE and MAE values. On the other hand, removing MFCC causes a comparatively smaller increase in RMSE and MAE, indicating its relatively lower contribution to the model's performance. It is noteworthy that each modality plays a distinct role, and their removal affects the model's predictive capabilities differently.

5.4. Depression Classification. In addition to estimating depression severity scores, we further conducted a classification task to distinguish between normal subjects and individuals with depression. This binary classification task allows us to evaluate the model's capability to differentiate between the two categories based on the threshold of 10 points on the Patient Health Questionnaire (PHQ-8) score [21]. To demonstrate the model's generalization, we also conducted the same testing on the AVEC-2014 [48] dataset which contains 150 subjects. The AVEC-2014 dataset utilizes BDI-II scores as labels, with a threshold of 21 to distinguish between individuals with depression and those without depression [49].

To provide a visual representation of our classification model's performance, we constructed a confusion matrix. The confusion matrix displays the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. This matrix provides insights into the model's strengths and weaknesses in terms of correctly and incorrectly classified instances which have been shown in Figure 6. 61 of 86 (70.93%) subjects with depression and 157 of 189 (83.07%) normal subjects were correctly predicted in E-DAIC dataset (Figure 6(a)), and 32 of 45 (71.11%) subjects with depression and 81 of 105 (77.14%) normal subjects were correctly predicted in AVEC-2014 dataset (Figure 6(b)).

For assessing the performance of our classification model, we employed the following evaluation metrics which can be seen from Figure 7. Accuracy (ACC) is defined as the proportion of correctly classified instances among all instances. Sensitivity (SEN) denotes the ratio of true positive predictions to the actual positive instances (depressed individuals). Specificity (SPE) is the ratio of true negative predictions to actual negative instances (normal individuals). Positive predictive value (PPV) is the proportion of true positive predictions among the instances that the model classified as positive. Negative predictive value (NPV) indicates the proportion of true negative predictions among the instances that the model classified as negative. Here, ACC, SEN, SPE, PPV, and NPV achieve 0.79, 0.71, 0.83, 0.66, and 0.86, respectively, in the E-DAIC dataset (Figure 7(a)) and the same index achieves 0.75, 0.58, 0.85, 0.71, and 0.78, respectively, in AVEC-2014 dataset (Figure 7(b)).

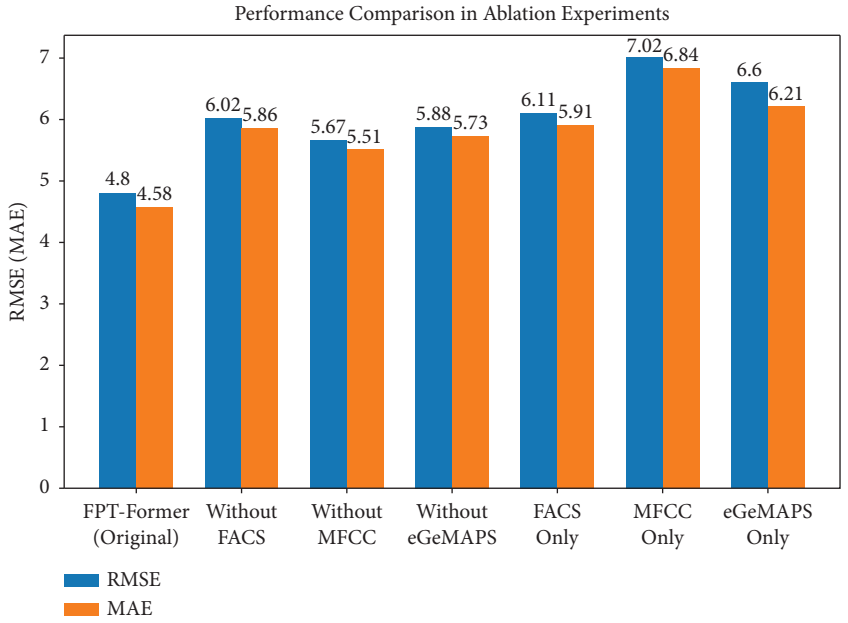

Figure 5: The effect of modality ablation on model performance.

Table 6: Comparison between FPT-Former (when only one modality is used) and baseline of AVEC-2019.

Model	RMSE
AVEC 2019 baseline-FACS [19]	7.02
AVEC 2019 baseline-MFCC [19]	7.28
AVEC 2019 baseline-eGeMAPS [19]	7.78
FPT-Former (FACS only)	**6.11**
FPT-Former (MFCC only)	**7.02**
FPT-Former (eGeMAPS only)	**6.60**

The bold font indicates that the RMSE of our model is lower than the previous three models.

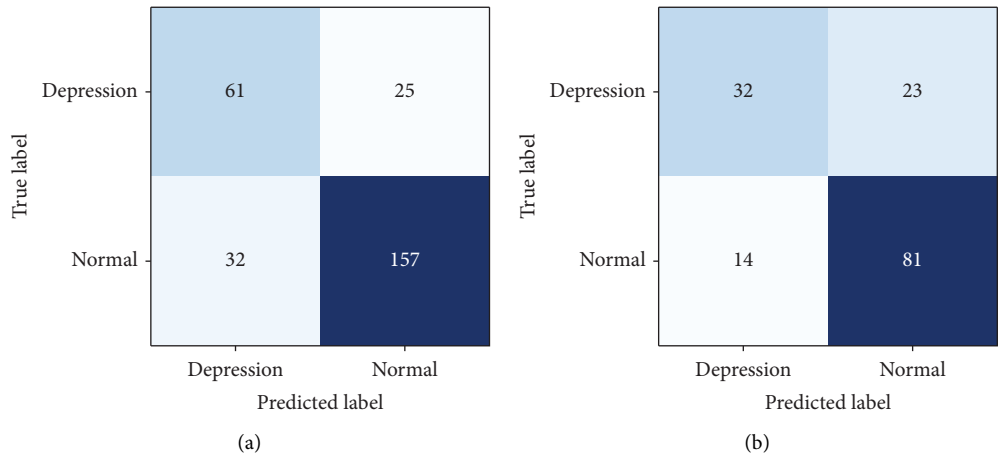

Figure 6: The confusion matrix of the classification result by E-DAIC dataset (a) and AVEC-2014 dataset (b).

5.5. Limitations and Future Works. Despite the promising results achieved by our proposed FPT-Former model, there exist several limitations that highlight avenues for future research.

First, the model is trained and evaluated using the E-DAIC dataset and AVEC-2014 dataset. Although these datasets are widely accepted, the generalizability of the model can be further validated using other multimodal datasets that are more diverse in terms of demographic characteristics and cultural contexts. Future work can involve conducting experiments on more datasets to improve the robustness and universality of the model.

Second, our study focused on three modalities: facial expressions, audio-MFCC, and audio-eGeMAPs. While these are undoubtedly important, depression manifests in

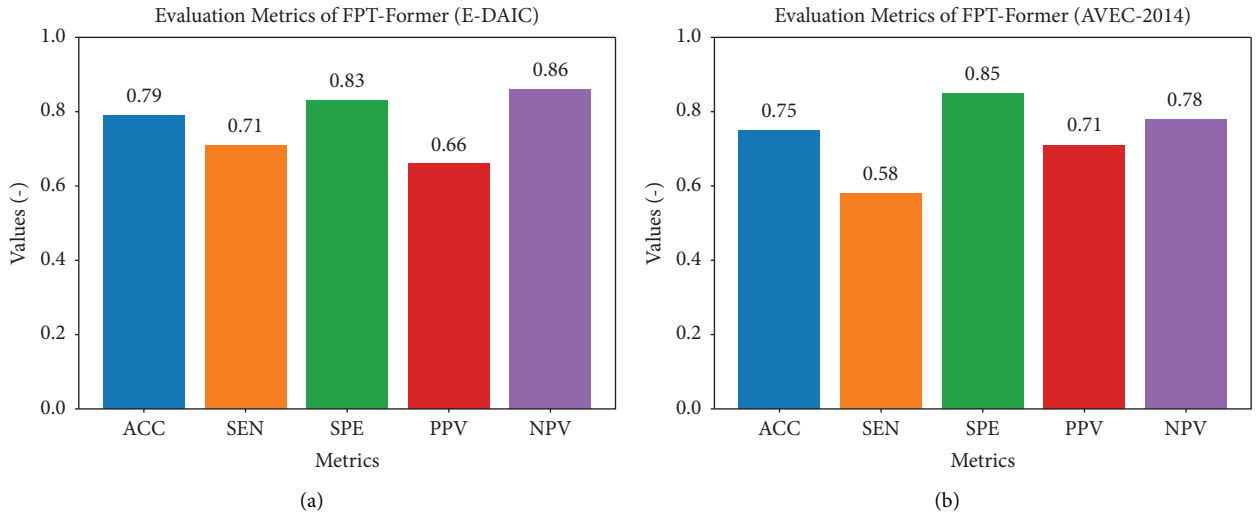

Figure 7: Evaluation metrics of FPT-Former classification performance.

various other ways. Future studies could consider incorporating additional modalities, such as text from patient interviews, physiological signals like heart rate variability, or even social interaction patterns [50].

Furthermore, it is essential to note that the current implementation of the FPT-Former relies on a concatenation method for multimodal fusion. In future research, we plan to explore more fusion techniques, such as attention mechanisms, tensor fusion, or hierarchical fusion, to enhance the model's accuracy and better capture the interdependencies among different modalities.

Despite these limitations, the FPT-Former represents a significant step towards a more comprehensive, accurate, and nuanced approach to depression severity estimation. Future work guided by these identified areas of improvement holds the potential to enhance the predictive capability of the model and broaden its applicability in real-world scenarios.

6. Conclusions

In this study, we introduced a novel flexible parallel transformer model, the FPT-Former, designed to harness the power of multimodal data in recognizing depression. Through its unique architecture, this model circumvents the challenge of quantitative differences across various modality features and provides a robust solution to reduce prediction errors. Besides, this model's ability to adapt to different numbers of measures across diverse modalities underlines its flexibility and applicability. Our FPT-Former model incorporates expert-knowledge-based audiovisual measures, facilitating the extraction of meaningful patterns from data, while maintaining the low dimensionality of input features. By employing low-dimensional measures as inputs, our model not only increases predictive efficiency but also addresses concerns related to personal privacy leakage, which is paramount in mental health applications. Experimental results on the E-DAIC dataset demonstrate the superiority of our model over existing techniques in terms of RMSE and MAE. The ablation studies further reveal the integral role each modality plays in achieving superior performance. Integrating multiple modalities and capturing long-term temporal dependencies from videos has the potential to detect depression accurately. After the comprehensive evaluation, the FPT-Former may become a useful diagnostic tool for mental health and contribute to global efforts in addressing this critical mental health issue.

In conclusion, this research contributes significantly to the understanding and technology of mental health diagnostics. The FPT-Former model, with its emphasis on expert-knowledge integration and privacy protection, not only advances the field of depression detection but also promotes the development of intelligent systems in mental healthcare. Its flexible structure and high predictive efficiency make it a potential tool for clinicians and researchers.

Data Availability

The data of the Extended Distress Analysis Interview Corpus (E-DAIC) can be applied and accessed at https://dcapswoz.ict.usc.edu/.

Additional Points

Highlights. (i) A flexible parallel transformer model has been proposed to recognize depression. (ii) Audiovisual expert-knowledge-based multimodal measures are integrated. (iii) Using low-dimensional measures as inputs increases predictive efficiency. (iv) The paralleled structure adapts different numbers of measures in diverse modalities. (v) The model achieves a comparable performance to the state-of-the-art works. (vi) Using expert-knowledge-based measures avoids personal privacy leakage.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

All authors contributed to the preparation of this manuscript. Yifu Li proposed the methodology, performed the experiments, and wrote the manuscript. Xueping Yang, Meng Zhao, and Zihao Wang proposed the methodology, wrote the manuscript, and performed statistical analysis. Yudong Yao, Wei Qian, and Shouliang Qi designed the study and proofread the manuscript.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (82072008), the Fundamental Research Funds for the Central Universities (N2224001-10), and open funding from Shenzhen Jingmei Health Technology Company Ltd.

References

[1] V. Patel, S. Saxena, C. Lund et al., "The Lancet Commission on global mental health and sustainable development," *The Lancet*, vol. 392, no. 10157, pp. 1553–1598, 2018.

[2] A. T. Beck and B. A. Alford, *Depression: Causes and Treatment*, University of Pennsylvania Press, Philadelphia, PA, USA, 2009.

[3] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Medicine*, vol. 3, no. 11, 2006.

[4] R. C. Kessler, P. Berglund, O. Demler et al., "The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R)," *Journal of the American Medical Association*, vol. 289, no. 23, pp. 3095–3105, 2003.

[5] D. D. Luxton, *Artificial Intelligence in Behavioral and Mental Health Care*, Elsevier Science, Amsterdam, Netherlands, 2015.

[6] A. B. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," *Psychological Medicine*, vol. 49, no. 9, pp. 1426–1448, 2019.

[7] A. Pampouchidou, P. G. Simos, K. Marias et al., "Automatic assessment of depression based on visual cues: a systematic review," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 445–470, 2019.

[8] S. Dhelim, L. Chen, H. Ning, and C. Nugent, "Artificial intelligence for suicide assessment using Audiovisual Cues: a review," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5591–5618, 2023.

[9] X. Zhou, Z. Wei, M. Xu, S. Qu, and G. Guo, "Facial depression recognition by deep joint label distribution and metric learning," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1605–1618, 2022.

[10] L. He, J. C.-W. Chan, and Z. Wang, "Automatic depression recognition using CNN with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, 2021.

[11] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2023.

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

[13] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology and Nonverbal Behavior*, 1978.

[14] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[15] F. Eyben, K. R. Scherer, B. W. Schuller 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, 2016.

[16] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," in *Proceedings of the 2010 4th International Conference on Signal Processing and Communication Systems*, pp. 1–5, IEEE, Gold Coast, Australia, December 2010.

[17] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: an emotional audio-textual corpus and a GRU/BiLSTM-based model," in *Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6247–6251, IEEE, Singapore, May 2022.

[18] H. Cai, Y. Gao, S. Sun et al., "Modma dataset: a multi-modal open dataset for mental-disorder analysis," 2020, https://arxiv.org/abs/2002.09283.

[19] F. Ringeval, B. Schuller, M. Valstar et al., "AVEC 2019 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*, pp. 3–12, Nice, France, October 2019.

[20] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[21] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[22] M. Du, S. Liu, T. Wang et al., "Depression recognition using a proposed speech chain model fusing speech production and perception features," *Journal of Affective Disorders*, vol. 323, pp. 299–308, 2023.

[23] W. Yang, J. Liu, P. Cao et al., "Attention guided learnable time-domain filterbanks for speech depression detection," *Neural Networks*, vol. 165, pp. 135–149, 2023.

[24] M. Niu, L. He, Y. Li, and B. Liu, "Depressioner: facial dynamic representation for automatic depression level prediction," *Expert Systems with Applications*, vol. 204, Article ID 117512, 2022.

[25] S. Kakuba, A. Poulose, and D. S. Han, "Deep learning-based speech emotion recognition using multi-level fusion of concurrent features," *IEEE Access*, vol. 10, pp. 125538–125551, 2022.

[26] S. Kakuba, A. Poulose, and D. S. Han, "Attention-based multi-learning approach for speech emotion recognition with dilated convolution," *IEEE Access*, vol. 10, pp. 122302–122313, 2022.

[27] Z. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, "MHA: a multimodal hierarchical attention model for depression detection in social media," *Health Information Science and Systems*, vol. 11, no. 1, p. 6, 2023.

[28] Y. Pan, Y. Shang, Z. Shao, T. Liu, G. Guo, and H. Ding, "Integrating deep facial priors into landmarks for privacy

preserving multimodal depression recognition," *IEEE Transactions on Affective Computing*, pp. 1–8, 2023.

[29] B. Zou, J. Han, Y. Wang et al., "Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2823–2838, 2023.

[30] L. Zhang, J. Zhao, L. He, J. Jia, and X. Meng, "An improved global-local fusion network for depression detection telemedicine framework," *IEEE Internet of Things Journal*, vol. 10, no. 22, pp. 20230–20240, 2023.

[31] J. Gratch, R. Artstein, G. M. Lucas et al., *The Distress Analysis Interview Corpus of Human and Computer Interviews*, LREC, Reykjavik, Iceland, 2014.

[32] J. Zhao, L. Zhang, Y. Cui, J. Shi, and L. He, "A novel Image-Data-Driven and Frequency-Based method for depression detection," *Biomedical Signal Processing and Control*, vol. 86, Article ID 105248, 2023.

[33] L. He, C. Guo, P. Tiwari, R. Su, H. M. Pandey, and W. Dang, "DepNet: An automated industrial intelligent system using deep learning for video-based depression analysis," *International Journal of Intelligent Systems*, vol. 37, no. 7, 2022.

[34] A. Singh, K. Raj, T. Kumar, S. Verma, and A. M. Roy, "Deep learning-based cost-effective and responsive robot for autism treatment," *Drones*, vol. 7, no. 2, p. 81, 2023.

[35] B. Jiang, S. Chen, B. Wang, and B. Luo, "MGLNN: semi-supervised learning via multiple graph cooperative learning neural networks," *Neural Networks*, vol. 153, pp. 204–214, 2022.

[36] A. M. Roy and J. Bhaduri, "DenseSPH-YOLOv5: an automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism," *Advanced Engineering Informatics*, vol. 56, Article ID 102007, 2023.

[37] S. Jamil and A. M. Roy, "An efficient and robust phonocardiography (pcg)-based valvular heart diseases (vhd) detection framework using vision transformer (vit)," *Computers in Biology and Medicine*, vol. 158, Article ID 106734, 2023.

[38] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: facial behavior analysis toolkit," in *Proceedings of the 2018 13th IEEE international conference on automatic face and gesture recognition (FG 2018)*, pp. 59–66, IEEE, Xi'an, China, May 2018.

[39] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[40] S. Memon, M. Lech, and L. He, "Using information theoretic vector quantization for inverted MFCC based speaker verification," in *Proceedings of the 2009 2nd International Conference on Computer, Control and Communication*, pp. 1–5, IEEE, Karachi, Pakistan, February 2009.

[41] R. Zilea, J. Navratil, and G. N. Ramaswamy, "Depitch and the role of fundamental frequency in speaker recognition," in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, IEEE, Hong Kong, China, April 2003.

[42] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, Firenze, Italy, October 2010.

[43] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, *Detecting Depression with Audio/Text Sequence Modeling of Interviews*, Interspeech, Dublin, Ireland, 2018.

[44] Y. Zhang, W. Hu, and Q. Wu, "Autoencoder based on cepstrum separation to detect depression from speech," in *Proceedings of the 3rd International Conference on Information Technologies and Electrical Engineering*, pp. 508–510, Hunan, China, December 2020.

[45] L. Yang, D. Jiang, and H. Sahli, "Feature augmenting networks for improving depression severity estimation from speech signals," *IEEE Access*, vol. 8, pp. 24033–24045, 2020.

[46] Z. Han, Y. Shang, Z. Shao et al., "Spatial-temporal feature network for speech-based depression recognition," *IEEE Transactions on Cognitive and Developmental Systems*, p. 1, 2023.

[47] M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu, "A multimodal fusion model with multi-level attention mechanism for depression detection," *Biomedical Signal Processing and Control*, vol. 82, Article ID 104561, 2023.

[48] M. Valstar, B. Schuller, K. Smith et al., "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pp. 3–10, Orlando, FL, USA, November 2014.

[49] K. L. Smarr and A. L. Keefer, "Measures of depression and depressive symptoms: beck depression Inventory-II (BDI-II), center for epidemiologic studies depression scale (CES-D), geriatric depression scale (GDS), hospital anxiety and depression scale (HADS), and patient health Questionnaire-9 (PHQ-9)," *Arthritis Care and Research*, vol. 63, pp. S454–S466, 2011.

[50] C. Lin, P. Hu, H. Su et al., "Sensemood: depression detection on social media," in *Proceedings of the 2020 international conference on multimedia retrieval*, pp. 407–411, Dublin, Ireland, June 2020.