informatics

MDPI

Article

Depression Detection Method Based on Multi-Modal Multi-Layer Collaborative Perception Attention Mechanism of Symmetric Structure

Shaorong Jiang [1,*,†], Chengjun Xu [2,3,*,†] and Xiuya Fang [1]

[1] School of Marxism, Jiaxing University, Jiaxing 314000, China; 00205921@stu.zjxu.edu.cn
[2] School of Artificial Intelligence, Jiangxi Normal University, Nanchang 330022, China
[3] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China
* Correspondence: maryjsr@163.com (S.J.); 005627@jxnu.edu.cn (C.X.)
† These authors contributed equally to this work.

Abstract

Depression is a mental illness with hidden characteristics that affects human physical and mental health. In severe cases, it may lead to suicidal behavior (for example, among college students and social groups). Therefore, it has attracted widespread attention. Scholars have developed numerous models and methods for depression detection. However, most of these methods focus on a single modality and do not consider the influence of gender on depression, while the existing models have limitations such as complex structures. To solve this problem, we propose a symmetric-structured, multi-modal, multi-layer cooperative perception model for depression detection that dynamically focuses on critical features. First, the double-branch symmetric structure of the proposed model is designed to account for gender-based variations in emotional factors. Second, we introduce a stacked multi-head attention (MHA) module and an interactive cross-attention module to comprehensively extract key features while suppressing irrelevant information. A bidirectional long short-term memory network (BiLSTM) module enhances depression detection accuracy. To verify the effectiveness and feasibility of the model, we conducted a series of experiments using the proposed method on the AVEC 2014 dataset. Compared with the most advanced HMTL-IMHAFF model, our model improves the accuracy by 0.0308. The results indicate that the proposed framework demonstrates superior performance.

Keywords: depression detection for college students; symmetric structure; attention mechanism; gender; emotional characteristics

Check for updates

Academic Editor: Jiang Bian

Received: 17 October 2025
Revised: 14 December 2025
Accepted: 8 January 2026
Published: 12 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Depression is a mental disorder that is prevalent globally. Patients with depression are usually in a state of persistent low mood, lose interest in daily activities, and suffer from problems such as lack of energy. These symptoms seriously affect their work and quality of life [1]. Globally, according to incomplete statistics, about 5% of adults suffer from depression to varying degrees [2]. Meanwhile, depression also threatens people's physical and mental health, and severely ill patients may have suicidal tendencies. As college students are an important group in society, their mental health deserves special attention. Statistics show that about 21.48% of college students in China suffer from depression [3]. Therefore, the early detection and intervention of depression is of great significance for promoting human health.

The existing depression detection methods are mainly divided into three categories: (1) traditional-based methods, (2) deep learning-based methods, and (3) depression detection methods based on the attention mechanism. Traditional methods mainly rely on fixed survey questionnaires (such as the PHQ-9) and the content of clinical interviews by psychiatrists. They are inefficient and easily affected by subjective factors. Subsequently, researchers propose using physiological signals for depression detection. For example, physiological indicators such as photoplethysmography (PPG), electrocardiogram (ECG), and electrodermal activity (EDA) are extracted through wearable devices to quantify emotional responses [4,5]. The depression level of individuals can be assessed by analyzing their voices. Traditional-based methods are often interfered with by subjective factors, resulting in possible biases and limitations in the diagnostic results, and they are not very efficient.

With the development of technology, methods based on deep learning have enhanced the objectivity of depression detection. Such methods are based on models such as the convolutional neural network (CNN) and long short-term memory (LSTM) to achieve deep representation learning of raw data. Marriwala et al. [6] proposed a hybrid architecture based on deep learning for depression detection using participants' audio and corresponding text transcriptions. Research shows that deep learning provides an efficient way for depression detection, and the accuracy of its text and audio models is as high as 0.9. Currently, research focuses on depression detection methods based on the attention mechanism. For example, Zhang et al. [7] used data such as images and texts to construct an attention-based multi-modal multi-task learning framework (AMM) for emotion recognition and depression detection. The results show that this model can make correct decisions using negative emotions and has good effects in emotion recognition and depression detection. Niu et al. [8] used audio and video data to detect individual depression levels by proposing a spatio-temporal attention network and a multi-modal attention feature fusion method. At the same time, research has begun to focus on the intrinsic relationship between depression and gender. Verma et al. [9] divided features into four categories based on gender and emotion to explore the influence of gender and emotion on depression recognition and used CNN and LSTM for depression recognition. Experiments showed that the gender-dependent models had better discriminative performance. Generally speaking, current research presents three major trends: data-driven multi-modal fusion for depression detection, embedding of gender and attention mechanisms, and using end-to-end deep learning frameworks for feature learning.

Although the above models have achieved some success, they have following deficiencies:

1. Existing studies generally focus on single-modal analysis and lack the full utilization of multi-modal data, resulting in insufficient feature extraction.
2. Some existing attention mechanisms do not fully consider data from different modalities. Especially in depression detection, only single-modality data is considered, and feature information such as gender is not taken into account, resulting in certain limitations in detection accuracy.
3. The existing models have complex structures and are parameter-heavy, resulting in relatively weak computational performance [1].

To address these limitations, this study introduces a new depression recognition model (SMMCA) featuring a symmetric structure with multi-modal multi-layer collaborative perception attention. The research contributions are mainly reflected in the following:

1. A depression detection model using a symmetric structure multi-modal multi-layer collaborative perception attention mechanism is proposed. This model incorporates multi-modal data, including emotional and gender characteristics, to systematically investigate their differential impacts on depression.

2. A multi-head attention mechanism module based on multi-layer perception is constructed. An interactive attention mechanism module is introduced. It can enable the model to effectively focus on the dynamic evolution process of emotional states, establish a deep association between emotion, gender information, and depression features, and fully explore the relationship among emotion, gender, and depression, thereby obtaining more important depression features.
3. We adopt a symmetric parallel structure and a lightweight design, such as parallel dilated convolution and a parallel multi-layer perceptron multi-head attention mechanism. This reduces computational complexity and facilitates the effective capture of cross-modal information. We utilized the publicly accessible and challenging AVEC 2014 dataset for comprehensive testing, and a comparative analysis was made with the most advanced HMTL-IMHAFF model. The prediction accuracy increased by 0.0308, the F1-score reached 0.892, and the Kappa coefficient was 0.837.

2. Related Study

2.1. Traditional Depression Detection Methods

In traditional methods, the diagnosis of depression mainly relies on patients' subjective descriptions, doctors' clinical interviews, and the use of fixed psychological scales, such as the Beck Depression Inventory (BDI) [10] and the Reynolds Adolescent Depression Scale (RADS-2) [11]. These methods are inefficient. Subsequently, researchers have identified depression by analyzing acoustic features, such as patients' speech rate and intonation [12], and have also used wearable devices, such as electroencephalogram and skin electrical activity, to extract physiological signals [13,14] to assess the degree of depression. However, physiological sensing devices are invasive. To better conduct depression detection, researchers have used multi-modal data fusion based on traditional methods. For example, Zhao et al. [15] introduced spectral subtraction and adopted a multi-modal fusion algorithm of speech signals and facial images for depression diagnosis. This method usually relies on carefully designed hand-crafted features, and its representation ability is limited.

2.2. Depression Detection Method Based on Deep Learning

With the rise of deep learning technology, the paradigm of depression detection has been transformed. Early studies applied models such as CNN and LSTM to single-modality data. For example, Amanat et al. [16] and Wonkoblap et al. [17] identified depressive states from social media texts. Jazaery et al. [18] used a 3D CNN to extract visual sequences, all achieving better results than traditional methods. In multi-modality depression detection using deep learning, He et al. [19] used deep learning techniques to extract features from audio and video for automatic depression detection. Jan et al. [20] combined facial expressions with voice signals and integrated deep learning with traditional methods to obtain a comprehensive assessment. Existing research shows that deep learning is effective in processing single-modality depression data, but it still has obvious deficiencies in tasks such as deep fusion of multi-modality information and handling complex long-range dependencies. For example, CNN still has certain limitations in dealing with long-time series dependencies [21].

2.3. Depression Detection Method Based on Attention Mechanism

To solve the above problems, Transformer can dynamically model the global dependencies among elements in a sequence through its unique self-attention mechanism. It has made breakthroughs in multi-modal tasks and has become the focus of current depression detection research. Current research is advancing in two directions. One is to construct a multi-modal fusion architecture. For example, Fan et al. [22] proposed a multimodal

feature enhancement network based on Transformer, which integrates video, audio, and remote photoplethysmogram modalities to fuse behavioral, acoustic, and physiological information. Zhou et al. [23] designed a fusion attention model. Zhang et al. [24] used a hybrid fusion method that combines attention decision fusion and feature extraction fusion for multi-modal depression diagnosis. The other is to refine the model and improve its interpretability. For example, Mahayossanunt et al. [25] introduced an attention mechanism into LSTM to focus on specific facial features. Experiments on the AVEC 2014 dataset showed that the model could capture important features of depression, such as head-turning and lack of smiling. Thekkekara et al. [26] constructed an attention-based CNN-BiLSTM model, emphasizing language features and so on, with an accuracy of 0.9671. In addition, researchers have also begun to introduce gender and emotion and build models through the attention mechanism, which shows that the research trend of depression detection has pointed toward personalized construction and model interpretability. Although these methods have made some progress, designing a lightweight and efficient model to achieve deep-level multi-modal fusion that is accurately applicable to key individual differences such as gender remains an important and urgent problem to be solved.

3. Methodology

3.1. Overview

As illustrated in Figure 1, the overall architecture of the model is composed of several key components, including feature extraction, a multi-layer perceptron (MLP) with multi-head attention (MHA) [27,28], and a BiLSTM module [29]. First, for a segment of input facial video data, a facial image was obtained through temporal sampling at fixed intervals. In the first branch, the facial emotion image $X \in R^{H \times W \times C}$ is input, where H and W denote the spatial dimensions of height and width, while C signifies the channel depth. Then, the feature F_1 is extracted from this image. In previous research, we found that there is a certain relationship between depression and gender. Therefore, different genders are considered in the second branch. The gender of the samples is used as a classification label, and Word2Vec is adopted to map these text labels into low-dimensional vectors, which serve as the initial semantic representation of gender. After that, the vector is further used in model training through a trainable embedding layer, and then discriminative gender features F_2 are obtained. Then, F_1 and F_2 are processed by the multi-layer perceptron multi-head attention mechanism to obtain new features F_1' and F_2'. Next, F_1' and F_2' are subjected to feature fusion through the interactive MHA to obtain a new feature, F_3. Subsequently, F_3 is further processed by the BiLSTM to get a new feature, F_3'. Finally, after global average pooling, it is input into the MLP classifier, and the Softmax activation function is used to map the features into a probability distribution to perform the binary classification task of depression vs. non-depression.

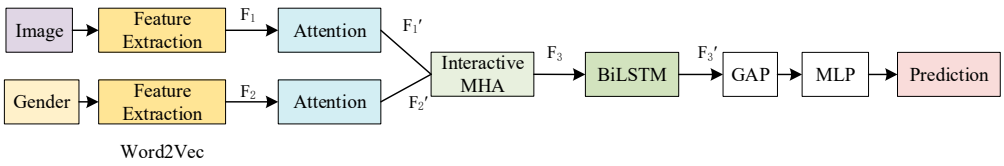

Figure 1. Model architecture. The key components are annotated as follows: (1) Facial Feature Extraction Branch. (2) Gender Feature Embedding Branch. (3) Multi-Layer Perceptron Multi-Head Attention Module (Attention). (4) Interactive Multi-Head Attention Module (Interactive MHA). (5) Bidirectional Long Short-Term Memory Network (BiLSTM). (6) Global Average Pooling (GAP). (7) Multi-Layer Perceptron (MLP). (8) Classifier. Among them, "image" represents a single image.

3.2. Feature Extraction Module

The architecture of the feature extraction module is depicted in Figure 2, and its formula is as follows:

$$B = SeLU(Conv2d(BN(X))) \tag{1}$$

$$Z_2 = SeLU(Conv2d(BN(Avg(B)))) \tag{2}$$

$$Z_3 = SeLU(Conv2d(BN(Z_2))) \tag{3}$$

$$Z_4 = SeLU(Conv2d(BN(Z_3))) \tag{4}$$

$$Z_4' = Z_4 + X \tag{5}$$

$$Z_5 = SeLU(Conv2d(BN(Z_4'))) \tag{6}$$

$$Z_6 = SeLU(Conv2d(BN(Avg(Z_5)))) \tag{7}$$

$$Z_7 = SeLU(Conv2d(BN(Avg(Z_6)))) \tag{8}$$

$$Z_7' = Z_7 + Z_4' \tag{9}$$

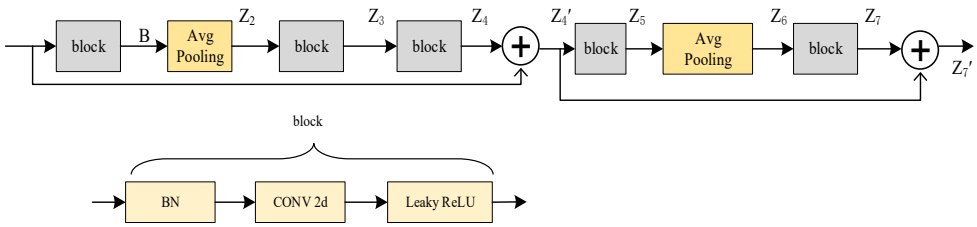

Figure 2. Feature extraction structure. The key components are annotated as follows: (1) Input and Initial Convolution Block. (2) First Average Pooling Layer. (3) Standard Convolution Block. (4) First Residual Connection. (5) Deep Convolution Block. (6) Second Average Pooling Layer. (7) Final Convolution Block. (8) Second Residual Connection.

Among them, Conv2d corresponds to a two-dimensional convolutional layer utilizing a kernel of dimensions 3×3, SeLU represents the Scaled Exponential Linear Unit, and Avg represents average pooling.

In the feature extraction module, the first branch and the second branch adopt the same structure. Taking the first branch as an example, first, the facial emotional image $X \in R^{H \times W \times C}$ is input. Before the convolution operation, batch normalization (BN) is used to effectively improve the convergence speed of the model and reduce the variance of gradient updates [30]. A Conv2d operation is then applied to the image, yielding a set of salient local features, as shown in Table 1. Finally, the SeLU activation function is applied. Its advantage lies in avoiding the gradient vanishing problem of SeLU in the negative value region and making the model more stable to obtain the feature B. Next, an average pooling operation is performed on B to decrease the spatial resolution of the features, lower the computational complexity, and at the same time retain important feature information,

obtaining Z_2. Then, after BN, Conv2d, and SeLU, Z_3 is obtained. Subsequently, with Z_3 as the input, continue the above-mentioned operation procedures of BN, Conv2d, and SeLU to obtain Z_4 in turn. After obtaining Z_4, as the model deepens, in order to prevent the disappearance of the original features and gradients, the residual connection is used. By establishing a skip connection linking Z_4 with the input facial emotion image X, we mitigate vanishing gradient effects and retain essential feature information, producing Z_4'.

Table 1. Differences between traditional convolution and parallel dilated convolution.

Methods	Kernel Size	Input Channel	Output Channel	Layer	Parameters	Total (M)
Ordinary	3×3	1024	1024	Conv1 Conv2 Conv3	$1024 \times 1024 \times 3 \times 3 = 9{,}437{,}184$ $1024 \times 1024 \times 3 \times 3 = 9{,}437{,}184$ $1024 \times 1024 \times 3 \times 3 = 9{,}437{,}184$	$2{,}381{,}155 \approx 23.8$
	5×5	1024	1024	Conv1 Conv2 Conv3	$1024 \times 1024 \times 5 \times 5 = 26{,}214{,}400$ $1024 \times 1024 \times 5 \times 5 = 26{,}214{,}400$ $1024 \times 1024 \times 5 \times 5 = 26{,}214{,}400$	$7{,}864{,}320 \approx 78.6$
Parrallel	7×7	512	512	Conv1 Conv2 Conv3	$512 \times 512 \times 7 \times 7 = 12{,}845{,}056$	$12{,}845{,}056 \approx 12.8$

Subsequently, taking Z_4' as the input, the batch normalization (BN), 2D Convolution (Conv2d), and SeLU operations are continuously performed to obtain Z_5. Then, Z_5 is subjected to the average pooling operation to obtain Z_6, which goes through the BN, Conv2d, and SeLU operations again to obtain Z_7. Finally, Z_7 is connected with Z_4' through a residual connection to obtain Z_7'. This enhances the hierarchical feature extraction ability of the model, enabling it to stably and efficiently capture key features and effectively overcome the deficiencies of the Swin-Transformer in processing fine-grained and highly continuous local patterns.

3.3. Multi-Layer Perceptron-Based Multi-Head Attention Mechanism Module

To effectively extract important features, a new multi-layer perceptron multi-head attention mechanism is proposed in this subsection, and its structure is shown in Figure 3.

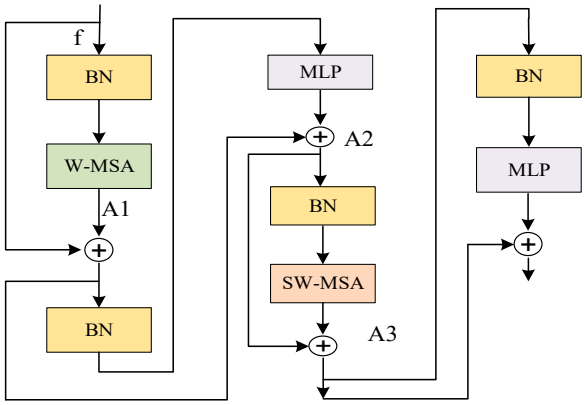

Figure 3. Structural diagram of the multi-head attention mechanism of the multi-layer perceptron. The key components are annotated as follows: (1) Window-based Multi-head Self-Attention (W-MSA). (2) Multi-Layer Perceptron (MLP). (3) Residual Connection (Add). (4) Shifted Window-based Multi-head Self-Attention (SW-MSA). (5) Multi-Layer Perceptron (MLP). (6) Residual Connection (Add).

The formula is formulated as follows:

$$A1 = f + \text{W-MSA}(\text{BN}(f)) \tag{10}$$

$$A2 = A1 + \text{MLP}(\text{BN}(A1)) \tag{11}$$

$$A3 = A2 + \text{SW-MSA}(\text{BN}(A2)) \tag{12}$$

$$\text{Atten} = A3 + \text{MLP}(\text{BN}(A3)) \tag{13}$$

Among them, f denotes the feature map obtained in the preceding phase and W-MSA represents the window-based multi-head self-attention mechanism [31]. MLP denotes multi-layer perceptron, while SW-MSA refers to the sliding window multi-head self-attention mechanism. Atten represents the finally obtained feature map.

In this module, first, the feature map f, obtained in the previous stage (as shown in Figure 3), is input. After batch normalization (BN) processing, it is input into the window-based multi-head self-attention mechanism (W-MSA) for feature extraction, and the feature A1 is obtained. This operation aims to capture the feature interactions within the local window through W-MSA, enhance the local representation ability of features, and retain the detailed information of features. Then, after A1 goes through BN processing, it is input into the MLP [32] for further feature extraction. After that, the input of the MLP is connected with A1 through a residual connection to obtain A2. In this operation, the MLP further enhances the feature expression ability, and the residual connection alleviates the gradient vanishing problem, making the model more stable and enabling it to obtain rich feature information. Subsequently, after A2 goes through BN processing, SW-MSA is used for feature extraction, and A3 is obtained. Here, SW-MSA enables information interaction between features in different windows through the shifted window approach, expanding the model's receptive field and enhancing its ability to capture global information [33]. Finally, after A3 goes through BN processing, it is input into the MLP again for feature extraction. Then, MLP's output is connected with A3 through a residual connection to obtain the final output Atten.

The above modules adopt W-MSA and SW-MSA and combine them with MLP and residual connections. This not only gradually enhances the feature representation ability, enabling it to have both local details and important global features, but also significantly enhances depression identification while maintaining computational efficiency, establishing a solid foundation for subsequent detection applications.

3.4. Interactive Multi-Head Attention Mechanism Module

On the basis of obtaining the relevant features F_1' and F_2', the interactive multi-head attention mechanism (interactive MHA) is adopted to effectively fuse important features [1]. Interactive MHA uses bidirectional attention to fuse features highly relevant to depression and emotions. It can effectively capture the complex relationships among emotions, gender, and depression, which helps to improve the features of depressive traits [1]. As illustrated in Figure 4, F_1' and F_2' are fused through interactive MHA to provide more abundant features for subsequent analysis.

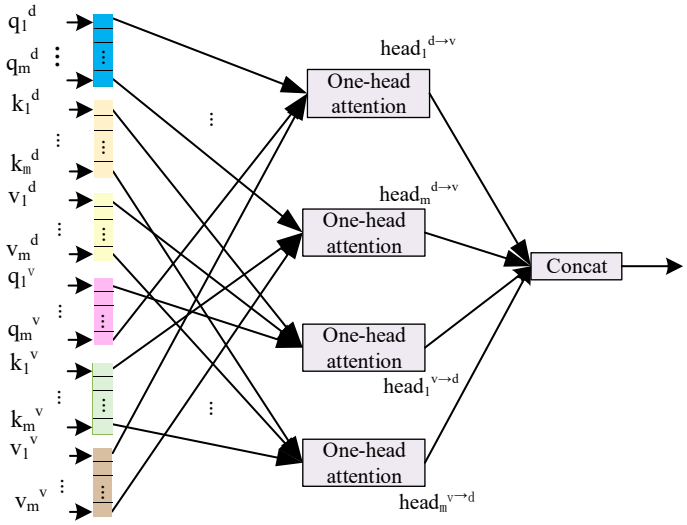

Figure 4. The IMHAFF framework, where q_1^d, k_1^d, v_1^d, q_1^v, k_1^v, and v_1^v represent the query, key, and value of depression-related features and emotion valence-related features, respectively.

3.5. BiLSTM Module

The feature F_3 is obtained through the above operations. To make accurate predictions using both previous and subsequent information and prevent information loss, the obtained feature F_3 is first flattened along the spatial dimension for subsequent temporal modeling. Then, it is fed into BiLSTM to extract local features and improve computational efficiency, as shown in Figure 5. Compared with the unidirectional LSTM, BiLSTM can obtain the complete left and right contexts at each position, taking into account the integrity of the context, enhancing the ability to model long-distance dependencies, and having stable gradients. It performs well in tasks such as depression detection and speech recognition [34,35]. In this study, by introducing the BiLSTM model, the relationship between the front and back of the sequence can be accurately captured to enhance the accuracy of depression identification. Subsequently, the Adam optimizer is incorporated, and learning rate scheduling is employed for optimization. The optimized results are then fed back to BiLSTM for iteration. After the iteration is completed, new features F_3' are obtained. This ensures the rapid convergence of the model and enhances its generalization ability. Finally, through global average pooling (GAP), the features output by BiLSTM are compressed into a fixed-length vector to retain global information. Then, the features F_3' are mapped layer by layer to a subspace with stronger discriminative power through a multi-layer MLP, thereby capturing fine-grained patterns related to depression, obtaining the probability of depression, and achieving the final depression prediction.

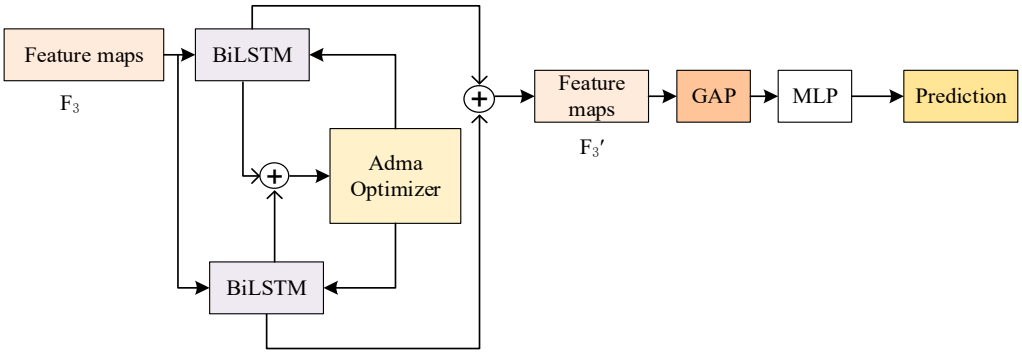

Figure 5. Structure diagram of BiLSTM.

In summary, F_3 flattens the sequence, divides it into blocks using a sliding window, and then inputs it into BiLSTM, effectively realizing the advantages of BiLSTM and enhancing the ability to model long-distance dependencies. At the same time, the Adam optimizer is also applied to embed the features to a lower-dimensional discriminative subspace highly correlated with depression, achieving efficient and accurate end-to-end prediction. On this basis, GAP and MLP work collaboratively to strengthen the representational capacity, enabling the model to more accurately identify depression features and improve the prediction accuracy.

4. Experiments

4.1. Dataset

To verify the feasibility and effectiveness of the model, we selected the AVEC 2014 dataset [36] for a large number of experiments. The AVEC 2014 dataset [36] is currently one of the representative public datasets used for depression recognition. Detailed information about the dataset is shown in Table 2.

Table 2. Detailed information about the dataset.

Dataset	Task Type	Training/ Testing Ratio	Sample Situation
AVEC 2014	Northwind: Participants read aloud a passage from the fable "The North Wind and the Sun."	80%/20%	Number of participants: 50. Original voice recordings: 50 (one per participant). Sample processing method: Each recording was preprocessed and segmented into fixed-length clips. Average video duration per session: About 25 min. Input clip duration: 3 s. Participant age range: 18–63 years Mean age ± SD: 31.5 ± 12.3 years.
	Freeform: Participants freely responded in German to a self-selected prompt, such as "What is your favorite dish?"	80%/20%	

This dataset reflects the participants' depression levels based on the Beck Depression Inventory-II (BDI-II) scores [37], shown in Table 3. When a participant's measured score was greater than or equal to 14, they were considered to be depressed. If the score was less than 14, the individual was considered normal. In addition, the AVEC 2014 dataset includes gender and other relevant sample information, as shown in Table 4.

Table 3. BDI-II scores and grades.

BDI Scores	Depression Level	Number of Videos	Valid Segments
0–13	Non-depressed	77	1435
14–19	Mild	22	411
20–28	Moderate	26	484
29–64	Severe	25	466

Table 4. Female/male sample distribution.

female	88	114	202
male	58	40	98
total	146	154	300

4.2. Experimental Environment and Evaluation Indicators

All experiments were performed under identical experimental conditions. As shown in Table 5, the Adam optimizer was employed with a learning rate of 10^{-4} and an exponential decay of 0.98 per epoch. The model underwent 50 training epochs using a batch size of 32. The video stream sampling rate was 15 fps, and the resolution was adjusted to 224 × 224.

To eliminate random interference, each experiment was repeated 10 times, and the results were recorded as mean ± standard deviation. The comparison models all adopted the official implementation configurations provided in the original papers.

Table 5. Experimental environment parameters.

Project	Content
Processor	Intel Xeon Gold 6248R @ 3.0 GHz (Santa Clara, CA, USA)
Memory	256 GB DDR4 ECC (Kingston: Fountain Valley, CA, USA—Headquarters)
Operating system	Ubuntu 20.04 LTS (Microsoft: Redmond, DC, USA)
Hard disk	2 TB NVMe SSD (RAID 0)(Western Digital: San Jose, CA, USA—Headquarters)
Software	Python 3.9.7 (MathWorks, Netik, MA, USA)
GPU	NVIDIA GeForce GTX 2080Ti (NVIDIA, Santa Clara, CA, USA)
Number of cycles	50
PyTorch	1.13.1 (Meta AI, Menlo Park, CA, USA)
Learning rate	10^{-4}
Training rate	5×10^{-5}
Momentum	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Weight decay	1×10^{-4}
Average pooling kernel size	2×2
Average pooling stride	2
padding	0
Number of filters in Conv layers	[64, 128, 256, 512]
Dropout rate	0.5
Stride in Conv layers	1
Weight initialization	He normal
Feature output dimensions	Varied per layer

The evaluation indicators included the confusion matrix, accuracy, precision, mean F1 score, Kappa index, and five-fold cross-validation.

4.3. Results and Analysis

The experimental findings are summarized in Table 6. Table 6 provides a comparison between the proposed model and the current most advanced models. We found that the SMMCA had superior performance in the depression detection task, as outlined below:

(1) Accuracy comparison. Previous research results have shown that deep learning methods have significant improvements compared to traditional methods [1]. Therefore, this paper compares against relatively new models. As shown in Table 6, when the training proportion was 80%, our model achieved an accuracy of 0.861. Compared with HMTL-IMHAFF [1], MMFAN [23], and STFN [38], the accuracy was improved by 0.0308, 0.047, and 0.052, respectively. As shown in Table 7, the prediction accuracy was 0.665 for males and 0.673 for females. Compared with HMTL-IMHAFF [1], the improvements were 0.0213 and 0.016, respectively. Further analysis shows that HMTL-IMHAFF [1] adopted the feature fusion method of interactive multi-head attention (IMHAFF) and a two-layer multi-task learning framework to analyze the intrinsic associations among emotion, gender, and depression. MMFAN [23] adopts a method that combines an attention model with multi-modal data input to extract facial and voice features for analyzing enhanced audiovisual sequence data to evaluate the degree of depression. Both of these methods enhance feature fusion and modeling capabilities through advanced architectures. HMTL-IMHAFF [1] uses an attention mechanism but fails to fully consider the relationships between different features, making it impossible to extract key features and limiting the accuracy of the model. Our model fully considers multi-modal data, enhances feature learning, and extracts

depression and gender features. Moreover, it also adopts the methods of multi-layer perceptron multi-head attention mechanism, interactive MHA, and BiLSTM, enabling the model to fully focus on the gender difference information between men and women, exploring the relationship among emotions, gender, and depression, and obtaining more important depression features. Therefore, the accuracy of our model is higher.

Table 6. Comparison of experimental results with the current most advanced models.

Model	Accuracy	F1-Score	Kappa
MMFAN [23]	0.814	0.798	0.731
HMTL-IMHAFF [1]	0.8302	0.8732	0.815
Bi-LSTM + CNN [34]	0.726	0.703	0.665
BERT-BiLSTM [39]	0.787	0.763	0.712
CNN + MFCC + spectrogram [40]	0.765	0.749	0.706
LSTM + MHA [41]	0.697	0.663	0.605
CNN + LSTM [9]	0.746	0.710	0.663
STFN [38]	0.809	0.781	0.725
Ours	0.861	0.892	0.837

Table 7. Comparison of accuracy for different genders.

Model	Accuracy	F1-Score	Female	Male
HMTL-IMHAFFH [1]	0.8302	0.7432	0.6570	0.6437
Ours	0.861	0.775	0.673	0.665

(2) The attention mechanism has been effectively verified. Further analysis shows that compared with MMFAN [23], the accuracy of HMTL-IMHAFF [1] increased by 0.0162. Our model improved by 0.0308 compared with HMTL-IMHAFF [1]. Through analysis, it was discovered that MMFAN [23] uses self-attention and channel attention mechanisms to extract the facial features of depression patients. However, this method lacks feature interaction between self-attention and channel attention. The HMTL-IMHAFF [1] model adopts an interactive multi-head attention mechanism, emphasizing the information interaction between multiple attention heads, thereby enhancing the representation and improving the generalization ability to comprehensively explore the in-depth relationships among gender, emotion, and depression and obtain an enhanced depression feature representation. Our proposed model not only adopts a multi-layer perceptron multi-head attention mechanism but also introduces interactive MHA. Through the collaboration of the two attention mechanisms, the model has significant advantages in capturing local and global features in depression feature extraction, which can significantly improve the accuracy. Therefore, different attention mechanisms have different impacts on the model. With a 0.0308 accuracy gain compared to the HMTL-IMHAFF model [1], the evidence strongly supports the effectiveness of our attention mechanism design.

(3) Reduction in the number of model parameters. Our method obtained an F1-score of 0.892 for the depression detection task, an increase of 0.0188 compared to the HMTL-IMHAFF model [1]. The improvement in the F1-score indicates that our model can effectively alleviate the bias caused by data imbalance in depression recognition. In addition, the Kappa coefficient of our model is 0.837, an increase of 0.022 compared to the HMTL-IMHAFF model [1], which reflects the degree to which the model's prediction results exceed random consistency and shows that its prediction results are more reliable. Further analysis reveals that the HMTL-

IMHAFF [1] model employs a traditional one-dimensional CNN for feature extraction and uses interactive multi-head attention (IMHA) for feature fusion. In contrast, the model we propose adopts a parameter-efficient symmetric structure, along with a co-design of the multi-layer perceptron multi-head attention mechanism and the BiLSTM module, to enhance the capture of key features. This not only reduces the number of parameters but also maintains strong representational ability. The experimental results show that through the innovative design of the symmetric structure, the multi-layer perceptron multi-head attention mechanism, and the BiLSTM module, our model achieves rapid convergence. Consequently, it outperforms existing models in both predictive precision and operational efficiency for depression assessment, providing a new technological paradigm for current depression detection.

4.4. Ablation Studies

To further verify the efficacy of different modules in the SMMCA model, we implemented three ablation experiments.

4.4.1. The Influence of Gender on Prediction Accuracy

To verify the influence of gender on model prediction, we conducted two groups of experiments: (1) experiments in which prediction was carried out using gender information; and (2) experiments in which prediction was carried out without using gender information. As shown in Table 8, the accuracy rate of prediction without using gender information was 0.835, and the accuracy rate of prediction using gender information was 0.861, with a decrease of 0.026 compared to the former. The experimental results show that gender difference information affects the extraction of depression-related features, and gender plays a role in depression prediction. It further verifies that there are differences in depressive symptoms and ways of expressing depression between females and males, which in turn impacts its discriminative effectiveness.

Table 8. The ablation test results.

Model	Accuracy	F1-Score
Without gender	0.835	0.856
Ours	0.861	0.892

4.4.2. Impact of the Attention Mechanism on Prediction Accuracy

To further verify the impact of the attention mechanism on model prediction, we conducted four groups of experiments: (1) the multi-layer perceptron multi-head attention mechanism was not adopted, but the interactive MHA mechanism was used; (2) the multi-layer perceptron multi-head attention mechanism was adopted, but the interactive MHA mechanism was not used; (3) neither the multi-layer perceptron multi-head attention mechanism nor the interactive MHA mechanism was adopted; and (4) both the multi-layer perceptron multi-head attention mechanism and the interactive MHA mechanism were used. As shown in Table 9, we found that the prediction accuracy decreased by 0.036 when the multi-layer perceptron multi-head attention mechanism was not used, indicating that the multi-layer perceptron multi-head attention mechanism plays an important role in the extraction of depression features. The prediction accuracy decreased by 0.048 when the interactive multi-head attention mechanism was not used, reflecting that this part plays an important role in feature interaction and information integration among features. The prediction accuracy was reduced by 0.098 when neither the multi-layer perceptron multi-head attention mechanism nor the interactive multi-head attention mechanism was used, while the prediction accuracy increased by 0.098 when both mechanisms were used,

indicating that the simultaneous use of the two attention mechanisms can significantly enhance the model's precision. The experimental results confirm that the attention mechanism plays an important role in depression prediction, further confirming the efficacy of the attention mechanism.

Table 9. The ablation test results.

Model	Accuracy	F1-Score
Without attention and with interactive MHA	0.825	0.837
With attention and without interactive MHA	0.813	0.825
Without attention and interactive MHA	0.763	0.781
Ours	0.861	0.892

4.4.3. Influence of BiLSTM on Prediction Accuracy

To further verify the impact of BiLSTM on model prediction, we conducted two groups of experiments: (1) prediction was carried out without using BiLSTM; (2) prediction was carried out using BiLSTM. As shown in Table 10, we found that the accuracy rate of prediction without using BiLSTM was 0.829, while the accuracy rate with BiLSTM was 0.861, and the accuracy rate decreased by 0.032 in comparison. The experimental results reveal that the BiLSTM module can successfully capture dependencies in sequential data and is useful for depression prediction, further verifying the important impact of BiLSTM on the accuracy of depression prediction.

Table 10. The ablation test results.

Model	Accuracy	F1-Score
Without BiLSTM	0.829	0.836
Ours	0.861	0.892

5. Conclusions

We proposed a new multi-modal multi-layer collaborative perception attention mechanism model based on a symmetric structure for depression detection. This model takes into account multi-modal feature information (such as gender and emotion) and uses parallel branches of a symmetric structure for feature extraction. We designed a multi-layer perceptron multi-head attention mechanism and introduced an interactive MHA to further explore the in-depth associations among emotion, gender, and depression, thereby obtaining more important depression features. To verify the effectiveness and feasibility of the model, we conducted a series of tests on the AVEC 2014 dataset. The experimental results demonstrate that compared with models such as HMTL-IMHAFF, MMFAN, and LSTM + MHA, the accuracy of our model reaches 0.861, exceeding these models by 0.0308, 0.047, and 0.164, respectively.

Although our model demonstrates clear superiority over other models, there is still room for further optimization of the model's complexity. In the future, we will endeavor to investigate the relationship between other factors and depression, such as age, to improve the prediction accuracy and reduce the model's computational burden, and apply it to the detection of depression in college students at various universities.

Author Contributions: Conceptualization, S.J. and C.X.; methodology, S.J. and C.X.; validation, X.F. and C.X.; formal analysis, S.J. and C.X.; writing—original draft preparation, S.J.; writing—review and editing, S.J., C.X. and X.F. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (42261068), the Natural Science Foundation of Jiangxi Province (20242BAB25112), the Industry-University-Research Collaborative Education Project of the Ministry of Education of China (220800247091048), and the Graduate Education Reform Project of Jiaxing University (651124009).

Data Availability Statement: The data associated with this research are available online. The AVEC2014 dataset is available for download at https://doi.org/10.1145/2661806.2661807.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AMM	Attentive multi-modal multi-task learning framework
CNN	Convolutional neural network
Conv	Convolution
BN	Batch normalization
LSTM	Long short-term memory
BiLSTM	Bidirectional long short-term memory
SMMCA	Symmetric structure multi-modal multi-layer collaborative perception attention
PHQ-9	Patient Health Questionnaire-9
PPG	Photoplethysmography
ECG	Electrocardiogram
EDA	Electrodermal activity
BDI	Beck Depression Inventory
RADS-2	Reynolds Adolescent Depression Scale, second edition
STA-DRN	Spatial–temporal attention depression recognition network
BERT	Bidirectional Encoder Representations from Transformers
MHA	Multi-head attention
SeLU	Scaled Exponential Linear Unit
Word2Vec	Word to Vector
Avg	Average pooling
MSA	Window multi-head self-attention
MLP	Multi-layer perceptron
W-MSA	window-based multi-head self-attention mechanism
SW-MSA	Shifted-window multi-head self-attention
GAP	Global average pooling
MMFAN	Multi-modal fused-attention network
STFN	The spatial–temporal feature network
MFCC	Mel-frequency cepstral coefficients
CNN-BiLSTM	Convolutional neural networks and bidirectional long short-term memory
HMTL-IMHAFF	Hierarchical multi-task learning framework based on interactive multi-head attention feature fusion

References

1. Xing, Y.; He, R.; Zhang, C.; Tan, P. Hierarchical Multi-Task Learning Based on Interactive Multi-Head Attention Feature Fusion for Speech Depression Recognition. *IEEE Access* **2025**, *13*, 51208–51219. [CrossRef]
2. Brookman, R.; Kalashnikova, M.; Conti, J.; Rattanasone, N.; Grant, K.; Demuth, K.; Burnham, D. Maternal depression affects infants' lexical processing abilities in the second year of life. *Brain Sci.* **2020**, *10*, 977. [CrossRef]
3. Luo, L.; Yuan, J.; Wu, C.; Wang, Y.; Zhu, R.; Xu, H.; Zhang, L.; Zhang, Z. Predictors of Depression among Chinese College Students: A Machine Learning Approach. *BMC Public Health* **2025**, *25*, 470. [CrossRef] [PubMed]
4. Giannakakis, G.; Grigoriadis, D.; Giannakaki, K.; Simantiraki, O.; Roniotes, A.; Tsiknakis, M. Review on psychological stress detection using biosignals. *IEEE Trans. Affect. Comput.* **2019**, *13*, 440–460. [CrossRef]
5. Schwartz, M.S.; Andrasik, F. *Biofeedback: A Practitioner's Guide*; Guilford Press: New York, NY, USA, 2017; pp. 68–113.
6. Marriwala, N.; Chaudhuri, D. Hybrid Model for Depression Detection Using Deep Learning. *Meas. Sens.* **2023**, *25*, 100587. [CrossRef]

7. Zhang, Y.; Li, X.; Rong, L.; Tiwari, P. Multi-task learning for jointly detecting depression and emotion. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021; pp. 3142–3149. [CrossRef]
8. Niu, M.; Tao, J.; Liu, B.; Huang, J.; Lian, Z. Multimodal spatiotemporal representation for automatic depression level detection. *IEEE Trans. Affect. Comput.* **2023**, *14*, 294–307. [CrossRef]
9. Verma, A.; Jain, P.; Kumar, T. An effective depression diagnostic system using speech signal analysis through deep learning methods. *Int. J. Artif. Intell. Tools* **2023**, *32*, 2340004. [CrossRef]
10. Von Glischinski, M.; von Brachel, R.; Hirschfeld, G. How "depressed" is "depressed"? A systematic review and diagnostic meta-analysis of the optimal cut-off points of the revised Beck Depression Inventory (BDI-II). *Qual. Life Res.* **2019**, *28*, 1111–1118. [CrossRef]
11. Ramos-Vera, C.; Quispe-Callo, G.; Bashualdo-Delgado, M.; Vallejos-Saldarriaga, J.; Santillán, J. Factorial and network structure ofthe Reynolds Adolescent Depression Scale (RADS-2) in Peruvian adolescents. *PLoS ONE* **2023**, *18*, e0286081. [CrossRef]
12. Kraepelin, E. *Manic-Depressive Insanity and Paranoia*; E & S Livingstone: London, UK, 1921; pp. 4–9.
13. He, Y.; Liang, F.; Wang, Y.; Wei, Y.; Ma, T. Advances in the Application of Wearable Devices in Depression Monitoring and Intervention. *Chin. J. Med. Devices* **2024**, *48*, 407–412. [CrossRef]
14. Li, M.; Li, J.; Chen, Y.; Hu, B. Detecting Stress Levels in College Students Using Affective Pulse Signals and Deep Learning. *IEEE Trans. Affect. Comput.* **2025**, *16*, 1942–1954. [CrossRef]
15. Zhao, J.; Su, W.; Jia, J. Depression Detection Algorithm Combining Prosody and Sparse Face Recognition. *Clust. Comput.* **2019**, *22*, 7873–7884. [CrossRef]
16. Amanat, A.; Rizwan, M.; Javed, A.R.; Alsaqour, R.; Pandya, S.; Uddin, M. Deep Learning for Depression Detection from Textual Data. *Electronics* **2022**, *11*, 676. [CrossRef]
17. Wongkoblap, A.; Vadillo, M.; Curcin, V. Depression Detection of Twitter Posters using Deep Learning with Anaphora Resolution: Algorithm Development and Validation. *JMIR Ment. Health*, 2021; *in press*. [CrossRef]
18. Al Jazaery, M.; Guo, G. Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Trans. Affect. Comput.* **2021**, *12*, 262–268. [CrossRef]
19. He, L.; Niu, M.; Tiwari, P.; Matin, P.; Su, R.; Jiang, J.; Guo, C.; Wang, H.; Ding, S.; Wang, Z.; et al. Deep Learning for Depression Recognition Using Audio-Visual Cues: A Review. *Inf. Fusion* **2022**, *80*, 56–86. [CrossRef]
20. Jan, A.; Meng, M.; Gaus, F.; Zhang, F. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Trans. Cognit. Develop. Syst.* **2018**, *10*, 668–680. [CrossRef]
21. Bhatt, D.; Patel, C.; Talsania, H.; Patel, J.; Vaghela, R.; Pandya, S.; Modi, K.; Ghayvat, H. CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope. *Electronics* **2021**, *10*, 2470. [CrossRef]
22. Fan, H.; Zhang, X.; Xu, Y.; Fang, J.; Zhang, S.; Zhao, X.; Yu, J. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. *Inf. Fusion* **2024**, *104*, 102161. [CrossRef]
23. Zhou, Y.; Yu, X.; Huang, Z.; Palati, F.; Zhao, Z.; He, Z. Multi-Modal Fusion Attention Network for Depression Level Recognition Based on Enhanced Audio-Visual Cues. *IEEE Access* **2025**, *13*, 37913–37923. [CrossRef]
24. Zhang, X.; Li, B.; Qi, G. A novel multimodal depression diagnosis approach utilizing a new hybrid fusion method. *Biomed. Signal Process. Control* **2024**, *96*, 106552. [CrossRef]
25. Mahayossanunt, Y.; Nupairoj, N.; Hemrungrojn, S.; Vateekul, P. Explainable depression detection based on facial expression u sing LSTM on attentional intermediate feature fusion with label smoothing. *Sensors* **2023**, *23*, 9402. [CrossRef]
26. Thekkekara, J.P.; Yongchareon, S.; Lesaputri, V. Attention-based CNN-BiLSTM model for depression detection from social media text. *Expert Syst. Appl.* **2024**, *249*, 123834. [CrossRef]
27. Botalb, A.; Moinuddin, M.; Al-Saggaf, U.M.; Ali, S.S.A. Contrasting Convolutional Neural Network (CNN) with Multi-Layer Perceptron (MLP) for Big Data Analysis. In Proceedings of the 2018 International Conference on Intelligent and Advanced System (ICIAS), Kuala Lumpur, Malaysia, 13–14 August 2018; pp. 1–5. [CrossRef]
28. AbdelRaouf, H.; Abouyoussef, M.; Ibrahem, M.I. An Innovative Approach for Human Activity Recognition Based on a Multi-Head Attention Mechanism. In Proceedings of the 2024 International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 18–20 December 2024; pp. 1559–1563. [CrossRef]
29. Hameed, Z.; Garcia-Zapirain, B. Sentiment Classification Using a Single-Layered BiLSTM Model. *IEEE Access* **2020**, *8*, 73992–74001. [CrossRef]
30. Xu, C.; Zhu, G.; Shu, J. A Combination of Lie Group Machine Learning and Deep Learning for Remote Sensing Scene Classification Using Multi-Layer Heterogeneous Feature Extraction and Fusion. *Remote Sens.* **2022**, *14*, 1445. [CrossRef]
31. Zhang, Y.; Li, T.; Li, C.; Zhou, X. A Novel Driver Distraction Detection Method Based on Masked Image Modeling for Self-Supervised Learning. *IEEE IoT J.* **2024**, *11*, 6056–6071. [CrossRef]

32. Desai, M.; Shah, M. Anatomy of Breast Cancer Detection and Diagnosis Using Multilayer Perceptron Neural Network (MLP) and Convolutional Neural Network (CNN). *Clin. Health Inform.* **2021**, *4*, 1–11. [CrossRef]
33. Xu, C.; Shu, J.; Zhu, G. Adversarial Remote Sensing Scene Classification Based on Lie Group Feature Learning. *Remote Sens.* **2023**, *15*, 914. [CrossRef]
34. Jo, A.-H.; Kwak, K.-C. Diagnosis of Depression Based on Four-Stream Model of Bi-LSTM and CNN From Audio and Text Information. *IEEE Access* **2022**, *10*, 134113–134135. [CrossRef]
35. Lin, L.; Chen, X.; Shen, Y.; Zhang, L. Towards automatic depression detection: A BiLSTM/1D CNN-based model. *Appl. Sci.* **2020**, *10*, 8701. [CrossRef]
36. Valstar, M.; Schuller, B.; Smith, K.; Almaev, T.; Eyben, F.; Krajewski, J.; Cowie, R.; AVEC, M.P. 2014: 3D dimensional affect anddepression recognition challenge. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Orlando, FL, USA, 7 November 2014; pp. 3–10. [CrossRef]
37. Niu, M.; Zhao, Z.; Tao, J.; Li, Y.; Schuller, B.W. Dual attention and element recalibration networks for automatic depression level prediction. *IEEE Trans. Affect. Comput.* **2022**, *14*, 1954–1965. [CrossRef]
38. Han, Z.; Shang, Y.; Shao, Z.; Liu, J.; Guo, G.; Liu, T.; Ding, H.; Hu, Q. Spatial–temporal feature network for speech-based depression recognition. *IEEE Trans. Cognit. Develop. Syst.* **2024**, *1*, 308–318. [CrossRef]
39. Cao, X.; Zakaria, L.Q. Integrating Bert With CNN and BiLSTM for Explainable Detection of Depression in Social Media Contents. *IEEE Access* **2024**, *12*, 161203–161212. [CrossRef]
40. Das, A.K.; Naskar, R. A deep learning model for depression detection based on MFCC and CNN generated spectrogram features. *Biomed. Signal Process. Control* **2024**, *90*, 105898. [CrossRef]
41. Zhao, Y.; Liang, Z.; Du, J.; Zhang, L.; Liu, C.; Zhao, L. Multi-head attention-based long short-term memory for depression detection from speech. *Front. Neurorobotics* **2021**, *15*, 684037. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.