



OPEN Depression detection methods based on multimodal fusion of voice and text

Zhenrong Xu^{1,2}, Yuan Gao^{1,2}, Fang Wang¹, Longqian Zhang¹, Li Zhang¹✉, Junke Wang¹ & Jie Shu¹

Depression is a prevalent mental health disorder, and early detection is crucial for timely intervention. Traditional diagnostics often rely on subjective judgments, leading to variability and inefficiency. This study proposes a fusion model for automated depression detection, leveraging bimodal data from voice and text. Wav2Vec 2.0 and BERT pre-trained models were utilized for feature extraction, while a multi-scale convolutional layer and Bi-LSTM network were employed for feature fusion and classification. Adaptive pooling was used to integrate features, enabling simultaneous depression classification and PHQ-8 severity estimation within a unified system. Experiments on the CMDC and DAIC datasets demonstrate the model's effectiveness. On CMDC, the F1 score improved by 0.0103 and 0.2017 compared to voice-only and text-only models, respectively, while RMSE decreased by 0.5186. On DAIC, the F1 score increased by 0.0645 and 0.2589, with RMSE reduced by 1.9901. These results highlight the proposed method's ability to capture and integrate multi-level information across modalities, significantly improving the accuracy and reliability of automated depression detection and severity prediction.

Keywords Depression, Deep learning, Multi-modal fusion, CMDC, DAIC-WOZ

Depression is a serious mental illness characterized by low mood, fatigue, loss of interest, and physical discomfort. It causes considerable suffering for patients and imposes a significant burden on socio-economic systems. According to estimates from the World Health Organization, approximately 350 million people worldwide suffer from depression, and this number continues to rise^{1,2}. Economically, the total burden of depression in the U.S. was estimated at \$326 billion in 2020³, which even exceeds Americans' expenditure on dining out⁴. Thus, enhancing the prevention, diagnosis, and treatment of depression has become increasingly urgent. Traditional diagnostic methods rely on clinical experience and self-assessment scales, however, subjects may conceal their true conditions, compromising the efficiency and reliability of detection. Therefore, developing a quick and accurate method for detecting depression remains a pressing challenge.

With the rapid advancement of artificial intelligence and machine learning technologies, depression detection systems utilizing multimodal data, including voice, text, and physiological signals, have increasingly become a research focus. These systems analyze individuals' behavioral patterns, linguistic expressions, and emotional states to assist clinicians in making timely and accurate assessments. Despite progress in multimodal depression detection research, extracting robust and highly discriminative features from individual modalities remains a significant challenge. In audio data feature extraction, most studies opt for eGeMAPS to extract voice features, however, its limited number of voice features leads to insufficient utilization of audio information⁵. Additionally, regarding feature fusion strategies, most studies focus primarily on simple feature concatenation or attention-based methods for integrating different modalities. However, simple concatenation fails to utilize features effectively. While attention mechanisms based on cross-modal interactions provide an alternative but come with challenges. They struggle to capture temporal dependencies, require large parameter sizes, and depend heavily on extensive training data. These factors together limit their effectiveness in depression detection⁶.

To address these challenges, this study proposes an innovative model that leverages Wav2Vec 2.0 for audio data and BERT for text data, forming an end-to-end architecture for automatic depression recognition. A multi-level aggregation layer is designed to effectively integrate cross-modal information while preserving temporal correlations between modalities. Additionally, a Bidirectional Long Short-Term Memory (Bi-LSTM) network is employed for classification, enabling the model to capture long-term dependencies in sequential data. Experiments were conducted on the authoritative Chinese Multimodal Depression Corpus (CMDC) and

¹School of Biomedical Engineering, South-Central Minzu University, Wuhan 430074, China. ²Zhenrong Xu, Yuan Gao contributed equally to this work. ✉email: zhangli1996@163.com

the English DAIC-WOZ datasets. The results demonstrate that the proposed fusion model performs effectively across different languages, highlighting its robust adaptability to diverse linguistic and contextual settings. This study presents a novel and effective approach for automated depression detection, contributing to the advancement of multimodal analysis in mental health diagnostics. By achieving consistent performance across different languages and datasets, the proposed method has the potential to support scalable, cross-linguistic applications in real-world clinical and telehealth environments.

Related work

Modalities related to depression

Previous studies have shown significant differences in various physiological data between depressed patients and healthy subjects during interviews^{7,8}. These differences are evident in both physiological and non-physiological data. Physiological data, such as heart rate variability (HRV), electrocardiogram (ECG), galvanic skin response (GSR), blood pressure, and electroencephalogram (EEG), have proven effective for depression recognition⁹. These indicators are strongly correlated with depression, and their measurement methods enable more accurate and reliable identification, reducing the potential for human concealment. Non-physiological data, by contrast, are easier to collect. Studies suggest that, compared to healthy individuals, depressed patients use more negative vocabulary, express pessimistic emotions, mention positive achievements or optimistic thoughts less frequently, and often make self-critical remarks¹⁰. Additionally, depressed patients typically have a monotonous voice, lacking variation, speak slowly with prolonged pauses, maintain low volume, and articulate unclearly. Their facial expressions show little variation, often appearing furrowed and vacant, with infrequent smiles¹¹. Due to the accessibility of non-physiological data, recognizing depression from this source has become a major research focus. However, research on multimodal depression detection is still limited. Therefore, this study uses both voice and text data for depression detection.

Research related to depression

In recent years, substantial progress has been made in depression recognition research focusing on single modalities, such as audio, video, and text. For example, Md. Rafiqul et al. collected user comments from Facebook and used Linguistic Inquiry and Word Count (LIWC) to analyze the text, assessing users' emotional states and sentence tenses based on word types to label their depressive tendencies. They extracted 27 features from sentences and classified depression using machine learning methods, achieving a maximum accuracy of 0.71¹². However, machine learning methods are limited in their ability to output results directly from raw data, and there may be issues with the data labeling, as some samples might not come from depressed patients. Amna et al. used a large, imbalanced dataset of tweets from Kaggle and proposed a depression prediction model based on deep learning, combining two layers of LSTM and two layers of dense RNN, achieving a prediction accuracy of 0.99¹³. However, the labels of such datasets may be controversial and may not serve as comprehensive depression recognition methods. Ahmed et al. developed an optimized word embedding method, applied it to Twitter user data, and developed a multi-channel depression recognition model by combining CNN and RNN, achieving an accuracy of 0.87¹⁴.

Additionally, some researchers have utilized audio data for depression recognition. Mamidiseti et al. collected audio data from 225 students in India, extracted voice features using openSMILE, and employed machine learning methods with classifiers such as Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), and K-Nearest Neighbors (KNN) for depression recognition, achieving an F1 score of 0.80¹⁵. Zhang et al. extracted voice features using a self-supervised model and combined them with transfer learning to identify depression, achieving a maximum recall rate of 0.96¹⁶. Although these studies have shown promising results, their reliance on a single modality limits the generalizability of the models.

Consequently, some researchers have started exploring multimodal data for depression detection. Multimodal data fusion is categorized into early and late fusion, with early fusion being favored by most researchers. Alhanai et al. extracted 379 features from both audio and text, concatenating them to input into SVM and LSTM networks for depression recognition, achieving a maximum F1 score of 0.77¹⁷. Lin et al. utilized a 1D Convolutional Neural Network and an LSTM network with a fused attention layer to process audio and text content separately, achieving a fusion of the two modalities through fully connected layers and obtained an F1 score of 0.85¹⁸. Zhang et al. designed a three-branch network to extract features from audio, text, and video, followed by feature fusion and attention modules for depression detection, achieving an F1 score of 0.78 and a recall rate of 0.81¹⁹. Ye et al. extracted 384 low-level voice features and proposed a multimodal fusion method based on deep spectral features and word vector features, achieving a detection accuracy of 0.91 and an F1 score of 0.90²⁰. However, in the aforementioned studies, only the research by Alhanai and Lin utilized professional depression datasets. The other datasets were derived from experimentally induced emotions in subjects, without involving actual patients diagnosed with depression or undergoing professional psychological evaluations. This limitation hinders an accurate assessment of detection performance in real cases of depression.

Zou et al. established a Chinese dataset for depression recognition, with the data verified by professional clinicians. They extracted text features using the BERT self-supervised model, video features via OpenFace, and audio features using eGeMAPS. Through early fusion for multimodal feature integration and processing for downstream tasks, they achieved a depression recognition F1 score of 0.94 and a recall rate of 0.91⁶. However, the limited audio features extracted using eGeMAPS constrain the full potential of end-to-end architectures, somewhat limiting model performance. Our previous research utilized the Wav2vec2.0 self-supervised model to extract voice features, applying a small fine-tuning network as a classification model, achieving an accuracy of 0.96²¹. However, as it only focused on a single modality in the English dataset, the model's generalization capability still requires enhancement. Building on this, we propose a new fusion model that uses BERT and Wav2vec2.0 to extract features from text and voice modalities, respectively. By further expanding the model's

receptive field through multi-scale convolutional kernels and integrating multimodal features using a mid-level fusion strategy, depression recognition is ultimately achieved based on the Bi-LSTM framework. Evaluation results on datasets from two different contexts demonstrate that the model exhibits significantly improved performance compared to the single-modality model.

Methods

Feature extraction

The voice of depressed patients often reflects their emotional and psychological states. This study employs Wav2vec 2.0 to extract features from audio data. Wav2vec 2.0 learns rich acoustic representations through extensive training on unlabeled data, producing high-dimensional features beneficial for emotion analysis tasks²². We used the Base version of Wav2vec 2.0 for voice feature extraction, taking the output from the final hidden layer, which yields a feature vector with 1024 dimensions. To ensure consistency, the temporal length of all voice features was standardized.

Text features are extracted using BERT. The contextual understanding architecture of BERT enables the model to generate rich and accurate semantic representations. Pre-training on a large corpus of unlabeled text, BERT demonstrates strong comprehension capabilities in sentence-level tasks. BERT is used for text feature extraction to generate deep language representations²³. The extracted features have a dimensionality of 768, matching the padded length of the sentences.

Feature fusion

Current multimodal fusion approaches primarily employ either feature concatenation or attention-based mechanisms. Several notable attention-based approaches have emerged in recent literature, including Yang et al.'s Multi-Scale Context layer that utilizes attention mechanisms to establish information flow across different scales through fully connected layers²⁴. Han et al.'s bidirectional fusion method employing sigmoid channel attention for quality-aware feature summation²⁵. Sun et al. developed the SimAM2 framework that innovatively incorporates principles from signal processing and uncertainty theory to enhance multimodal data integration²⁶. Similarly, Vaswani et al. proposed a Transformer-based architecture that effectively enables bidirectional cross-modal attention mechanisms for feature fusion²⁷. However, the above attention-based methods typically demand substantial memory resources, often exceeding the capacity constraints of real-world clinical deployment environments. Moreover, when processing temporal sequences characteristic of speech and language data, pure attention architectures show reduced sensitivity to timing patterns compared to hybrid CNN-LSTM networks. To address this, we introduce a multi-scale convolution (MSC) module to enhance the feature representation, enabling the model to capture information at various scales. The multi-scale information is concatenated along the channel dimension. Then, adaptive average pooling is used to reduce the sequence length and align voice and text features into a shared semantic space. Finally, the fused features are input into the classification network. Different channels represent the text or speech features at specific time steps. The feature fusion process is illustrated in the feature fusion section of Fig. 1.

Fusion model architecture

Figure 1 illustrates the depression recognition framework based on voice and text bimodal data. First, the text and audio features are extracted separately using BERT and Wav2Vec2.0. The extracted features are then fed into a multi-scale convolutional layer, where convolution kernels of varying sizes capture information across different time scales, the methods of calculation are as follows:

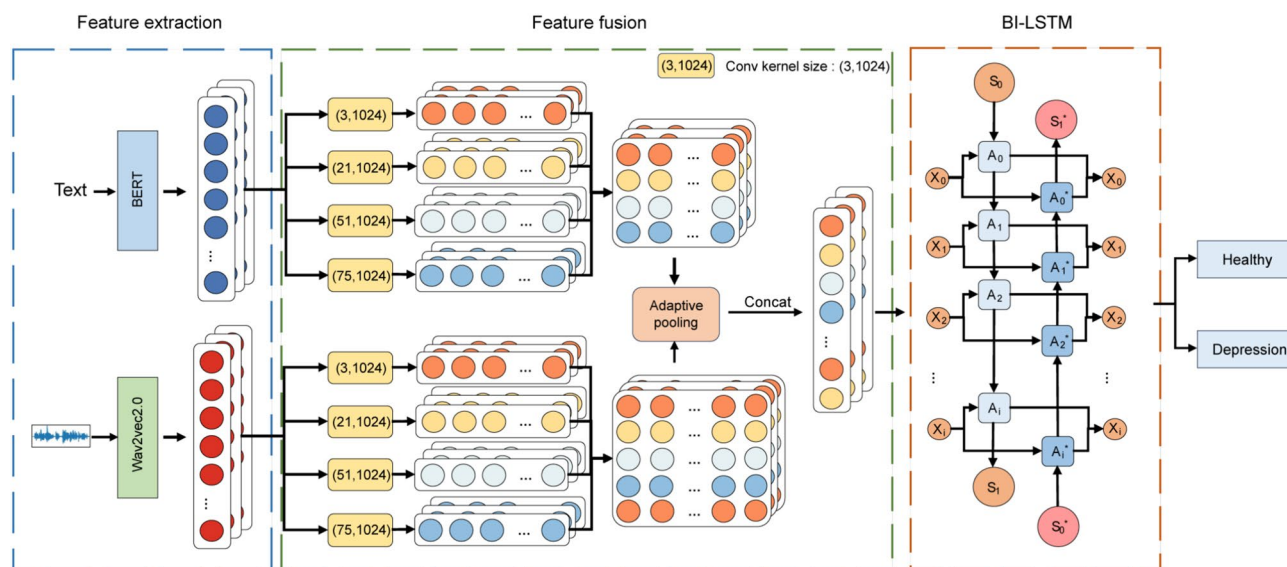


Fig. 1. Model architecture

$$f_{kernel}[m, n] = \sum_j \sum_k o[j, k] f[m - j, n - k] \quad (1)$$

f_{kernel} denotes the features extracted using kernels of different scales, where m and n represent the sequence length and feature dimension of the feature matrix, o is the convolution kernel, and j and k represent the elements within the kernel. f is the input raw matrix. Through convolution operations, the model updates information across different data positions, capturing spatial information from text or audio modalities and thereby enhancing feature expressiveness.

After extracting the multi-scale convolution features f_{kernel} , the features from different scales of a single modality are concatenated in the channel dimension. To achieve adaptive output dimensions, adaptive average pooling was applied. Firstly, to determine the area Δ_i in the input tensor corresponding to the output position i' , it needs to calculate the ratio r of the entire input tensor length n to the target output length n' :

$$r = \frac{n}{n'} \quad (2)$$

Then, for output position i' , the length of the corresponding area Δ_i in the input tensor is calculated as follows:

$$\Delta_i = \frac{(i' + 1) \cdot r - i' \cdot r}{r} \quad (3)$$

Finally, the output text or audio feature matrix $Y_m^{i'}$ is calculated as follows:

$$Y_m^{i'} = \frac{1}{\Delta_i} \sum_{i=\alpha_{i'}}^{\beta_{i'}} F_m^i \quad (4)$$

where F_m^i represents the audio or text modality features extracted at time step i through the MSC module. $\alpha_{i'}$ is the starting index, and $\beta_{i'}$ is the ending index of the input region corresponding to the output position i' . During feature fusion, the audio and text feature matrices are concatenated and fed into the Bi-LSTM for context modeling.

Bi-LSTM uses a bidirectional gating mechanism to extract sequence information in both directions, thus enhancing the model's understanding of the intermodal context. The core of Bi-LSTM consists of three gating units (forget gate, input gate, and output gate), which control the retention and updating of information. The forget gate determines which long-term state information in the sequence should be retained, and its computation is as follows:

$$f_i = \sigma(W_f, [h_{i'-1}, Y_{i'}] + b_f) \quad (5)$$

where W_f and b_f represent the weight matrix and bias vector of the forget gate, σ denotes the sigmoid function, $h_{i'-1}$ represents the output of the hidden state at the previous time step, and $Y_{i'}$ denotes the input at the current time step.

Subsequently, the input gate is used to control how much new information can be updated to the current state. It decides which parts of the candidate state are added to the existing cell state. The specific calculation method is as follows:

$$t_i = \sigma(W_t, [h_{i'-1}, Y_{i'}] + b_t) \quad (6)$$

$$C'_i = \tanh(W_c, [h_{i'-1}, Y_{i'}] + b_c) \quad (7)$$

Where W_t, b_t, W_c , and b_c are the parameters related to the input gate, and the candidate state, and \tanh denotes the hyperbolic tangent activation function. Then, the current state is updated based on the results of the forget gate and the input gate. The calculation for the current input transformation matrix C_i is as follows:

$$C_i = f_i * C_{i-1} + t_i * C'_i \quad (8)$$

Here, $*$ denotes element-wise multiplication. The output gate controls the output at the current time step. The calculation formula is as follows:

$$O_i = \sigma(W_o, [h_{i'-1}, Y_{i'}] + b_o) * \tanh C_i \quad (9)$$

W_o and b_o represent the weight matrix and bias vector of the output gate. The output is calculated by element-wise multiplication of the non-linearly activated input transformation matrix and the element-wise sigmoid-transformed output proportion matrix, resulting in the hidden state at the current time step. When the model reaches the final time step, the hidden state is extracted as the forward iteration result. Simultaneously, the data from the final time step is used as input to obtain the backward hidden state. The hidden states from both directions are then concatenated. Finally, the features are mapped through a fully connected layer to output the classification results or PHQ-8 prediction scores, thereby achieving depression detection.

In the proposed multimodal architecture, pre-trained models are utilized to effectively extract features. Multi-level aggregation modules are employed to fuse features across different time scales, enhancing the accuracy of feature extraction. Adaptive average pooling is applied to compress the features, reducing dimensionality and ensuring effective utilization of information. Finally, Bi-LSTM is used to enhance cross-modal context understanding, leading to improved emotion classification and depression detection.

Experiments
Datasets

The CMDC Depression dataset was constructed based on semi-structural interviews in Chinese, designed to support the screening and assessment of severe depression⁶. All participants were recruited from a well-known psychiatric hospital in China, aged between 20 and 60 years, with diagnoses meeting the criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM). The interviews were designed by professional clinicians and conducted by trained interviewers. Each subject (19 severe depression patients and 26 healthy controls) has corresponding audio and text records, with the transcripts derived from the audio recordings and manually verified. The dataset includes PHQ-9 and HAMD-17 scores, offering robust data for the screening and evaluation of severe depression in China.

The DAIC-WOZ dataset was built using semi-structural interviews in English, conducted by a virtual human named Ellie. It forms part of a larger clinical interview corpus known as the Distress Analysis Interview Corpus (DAIC). This methodology supports the development of automated assessment of language and non-verbal behaviors associated with depression and Post-Traumatic Stress Disorder(PTSD). The dataset includes 189 subjects. PHQ-8 results indicate that 56 participants exhibit varying degrees of depressive symptoms. The dataset consists of text, video, and audio data related to depression²⁸, offering valuable resources for the automated detection of depression.

Our audio data preprocessing pipeline began with standardization of all recordings to 50-60 dB amplitude range, accompanied by noise reduction to enhance signal quality. To address dataset size limitations while preserving linguistic context, we implemented a semantic-aware segmentation approach where each sample contains five consecutive sentences, with detailed parameters documented in Table 1.

For multimodal alignment, we established a rigorous verification protocol. Transcripts were systematically cross-validated against their corresponding audio recordings, with any discrepancies corrected to ensure textual accuracy. This process identified and removed temporally misaligned segments, representing a minimal 1% data loss that did not compromise overall dataset integrity.

The validated audio clips were uniformly processed into 20-second segments using zero-padding when necessary, creating consistent input dimensions for downstream analysis. Following segmentation, each sample underwent manual quality inspection to identify and rectify any processing artifacts. This meticulous curation yielded final datasets with clinically meaningful sample sizes: the CMDC dataset contains 2969 samples (1677 healthy controls vs. 1292 depression cases) and the DAIC-WOZ dataset comprises 6544 samples (4515 healthy controls vs. 2029 depression cases).

Figure.2 illustrates this comprehensive preprocessing workflow, demonstrating the multilayered approach to data quality assurance. The resulting datasets maintain clinically relevant class distributions while eliminating technical artifacts that could potentially bias model performance. Importantly, our segmentation strategy successfully balanced the need for adequate sample sizes with preservation of meaningful linguistic units for depression detection.

Model training

Information about the experimental environment for this study is as follows: CPU: 11th Gen Intel(R) Core (TM) i7-11700 @ 2.50 GHz; GPU: NVIDIA GeForce RTX 3090; RAM: 24GB. Operating system: 64-bit Ubuntu 20.04.4 LTS; CUDA: 11.6; Python 3.7.

The model was trained for 100 epochs using an early stopping strategy. If no significant improvement in model performance occurred over a predefined number of epochs, the training was stopped early. Cross-validation was performed by partitioning the dataset into training, validation, and testing sets in a 6:2:2 ratio. After processing the voice data with Wav2vec 2.0, the resulting feature dimensionality was 1024, while the text feature dimensionality after processing with BERT was 768. Subsequently, 2D convolutional kernels with sizes (3/21/51/75, 1024) and (3/21/51/75, 768) were applied to expand the receptive fields of the voice and text data, respectively. A dropout rate of 0.2 was set uniformly for all convolutional modules. A feature mapping module was then utilized to compress the voice and text features, with the compression convolutional kernels set to (1, 3). After adaptive pooling, the audio data is compressed to a final padding length of 64. The training task determines the output size. For classification tasks, the output is 2, and for the regression task that aims to predict the PHQ-8 score, the output is 1. The specific parameters of each module are summarized in Table 2.

Parameter	Number of parameters
Minimum duration of silence	300 ms
Silence decibel threshold	− 45 dB
Effective voice length	100–10000 ms
Keep the silence length	300 ms

Table 1. Audio segmentation parameters

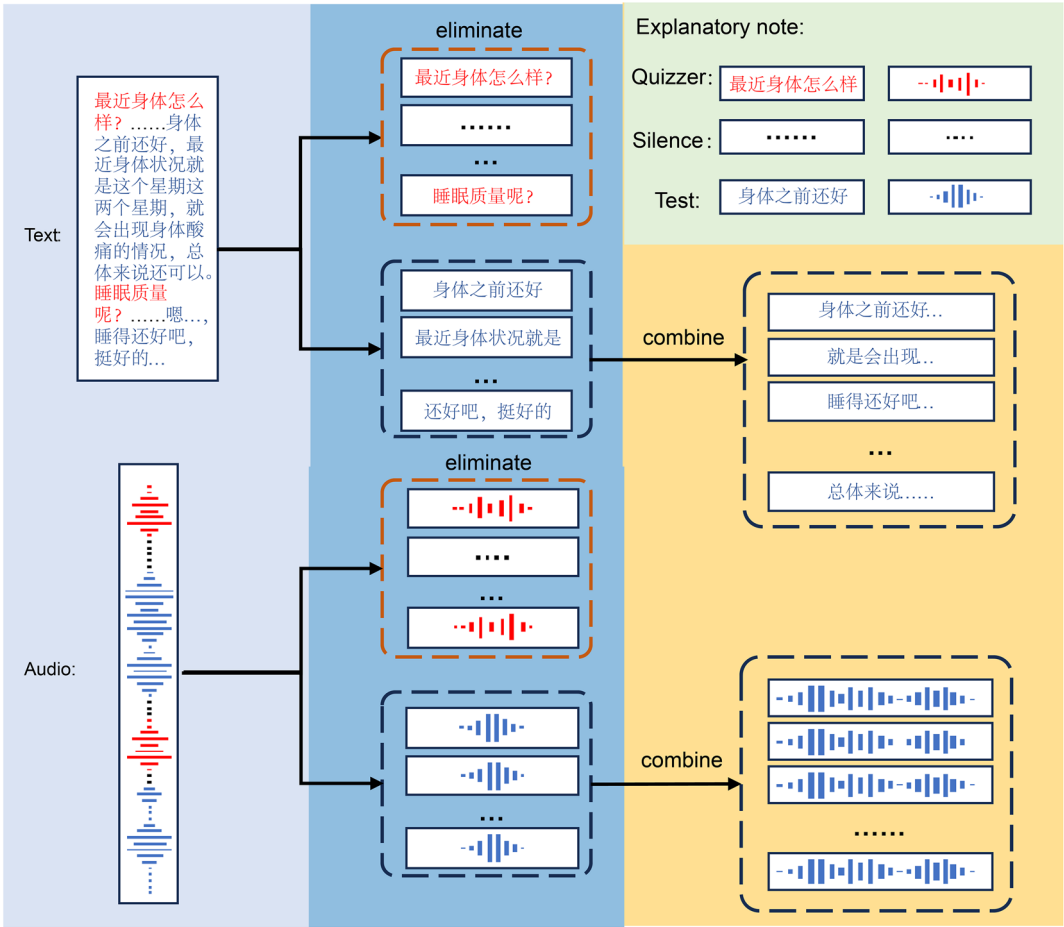


Fig. 2. The figure shows the data processing flow. Red represents the quizzer segments, black indicates silence segments, and blue denotes the subject's data segments.

Hierarchy	Hyper-parameter
MSC	Conv (3/21/51/75, 768/1024)
Feature fusion	Average pooling (length, 64)
Bi-LSTM	Number layers = 4, hidden size=128
Output size	Number class = 2 or 1

Table 2. Specific parameters of each module

Evaluation metrics

Despite preprocessing and data augmentation efforts to mitigate data imbalance, the final dataset ratios remain mildly imbalanced (CMDC: 56.5% healthy vs. 43.5% depressed; DAIC-WOZ: 69% healthy vs. 31% depressed). While these distributions reflect real-world clinical screening scenarios, the mild class imbalance necessitates comprehensive performance evaluation. The confusion matrix is a standard tool for assessing classification performance. It enables the calculation of various performance metrics, including accuracy, precision, recall, and the F1 score. The confusion matrix categorizes data into four categories: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), corresponding to correctly predicted positives, incorrectly predicted positives, incorrectly predicted negatives, and correctly predicted negatives, respectively. Accuracy is defined as the proportion of correctly classified samples, calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{10}$$

Precision refers to the ratio of true positives to the total number of samples predicted as positive:

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

Recall is the ratio of true positives to the total number of actual positive samples:

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

The F1-score (harmonic mean of precision and recall) is employed as the primary metric due to its robustness for imbalanced data. This approach ensures that both depression detection sensitivity and diagnostic precision are given balanced consideration. The specific calculation is as follows:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (13)$$

Each of these four metrics focuses on different aspects: accuracy considers all types of predictions, while precision and recall concentrate on specific prediction types. The F1 score balances precision and recall, making it particularly suitable for imbalanced datasets. To predict the severity of depression in samples, this study extends the model to a regression task aimed at estimating the PHQ-8 index. The predictive performance is evaluated using the MAE and RMSE metrics. MAE measures the average absolute difference between predicted and actual values, and its formula is as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (14)$$

RMSE considers the mean of squared errors, assigning higher weights to larger deviations. Its formula is expressed as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (15)$$

Here, n denotes the number of samples, \hat{y}_i represents the predicted values, and y_i indicates the actual values. MAE reflects the average deviation between predictions and true values. RMSE is more sensitive to larger errors. Together, these metrics provide a comprehensive evaluation of the model's predictive performance.

Results and discussion

The proposed fusion model integrates a MSC module and a Bi-LSTM module. To better understand the individual contributions of each module to the overall performance, we compared the test results of unimodal and multimodal data on the MSC model, the Bi-LSTM model, and the fusion model proposed in this study. Specifically, in the experimental design, the MSC model directly passed its output through a fully connected layer for final decision making. In contrast, the Bi-LSTM model compresses feature dimensions using a fully connected layer before making predictions. These ablation experiments provided valuable insights into the impact of different model architectures on prediction performance and highlighted the potential advantages of leveraging multimodal information. Furthermore, to visually illustrate the differences in features before and after fusion, we employed the t-Distributed Stochastic Neighbor Embedding (t-SNE) technique to perform a visualization analysis of unimodal and multimodal features. This analysis offered a clear representation of how feature fusion enhances the learning capability of the model. Finally, we compared the test results of the proposed fusion model with those of other state-of-the-art methods evaluated on the same datasets. This comprehensive comparison allowed us to identify the strengths and weaknesses of different approaches, providing a holistic evaluation of the proposed model's effectiveness and its position relative to existing methods.

Evaluation results

Our systematic evaluation of individual module components reveals distinct performance characteristics, as demonstrated by the CMDC dataset results in Table 3.

The ablation study highlights the complementary nature of each architectural element: The multi-scale convolution (MCN) network exhibits particular efficacy in multimodal data processing, attaining a 0.9673 F1-score through its hierarchical feature extraction mechanism. This advantage stems from the network's capacity to concurrently capture both granular acoustic characteristics and higher-level spatial patterns within combined audio-text feature spaces.

Conversely, the Bi-LSTM network demonstrates specialized proficiency in temporal sequence modeling, achieving exceptional unimodal audio performance (accuracy = 0.9613, F1-score = 0.9605). These results validate the architecture's competence in modeling temporal speech patterns and maintaining sequential dependencies. Nevertheless, the Bi-LSTM's effectiveness diminishes when processing concatenated multimodal features (F1-score = 0.9213), reflecting the challenges inherent in heterogeneous feature integration.

The integrated MCN+Bi-LSTM framework synthesizes these complementary strengths, yielding optimal performance metrics. The complete system shows measurable improvements over both constituent modules: a 0.35% F1-score enhancement (0.9708 vs. 0.9673) relative to the MCN network alone, coupled with a more

Modality	Metric	MCN	Bi-LSTM	MCN+Bi-LSTM
Audio	Accuracy	0.8636	0.9613	–
	Precision	0.8751	0.9608	–
	Recall	0.8772	0.9604	–
	F1 score	0.8636	0.9605	–
Text	Accuracy	0.7604	0.7710	–
	Precision	0.7577	0.7684	–
	Recall	0.7534	0.7722	–
	F1 score	0.7549	0.7691	–
Audio+Text	Accuracy	0.9680	0.9226	0.9747
	Precision	0.9691	0.9206	0.9752
	Recall	0.9658	0.9221	0.9664
	F1 score	0.9673	0.9213	0.9708

Table 3. Evaluation results of the CMDC dataset Significant values are in bold.

Modality	Metric	MCN	Bi-LSTM	MCN+Bi-LSTM
Audio	Accuracy	0.9434	0.9496	–
	Precision	0.8603	0.9608	–
	Recall	0.9277	0.9157	–
	F1 score	0.8928	0.9021	–
Text	Accuracy	0.7021	0.7471	–
	Precision	0.6506	0.6723	–
	Recall	0.7020	0.7470	–
	F1 score	0.6753	0.7077	–
Audio+Text	Accuracy	0.9343	0.9274	0.9794
	Precision	0.8170	0.7926	0.9702
	Recall	0.9548	0.9669	0.9631
	F1 score	0.8806	0.8711	0.9666

Table 4. Evaluation results of the DAIC-WOZ dataset Significant values are in bold.

substantial 2.67% gain (0.9708 vs. 0.9213) compared to the standalone Bi-LSTM implementation. These results substantiate the framework’s dual capability to both enhance the MCN’s multimodal processing through temporal refinement and substantially augment the Bi-LSTM’s capacity for heterogeneous feature analysis.

Similar to the results observed on the CMDC dataset, the proposed fusion model achieves superior performance on the DAIC-WOZ dataset, as shown in Table 4, further validating its robustness and effectiveness across different datasets and experimental settings. The method exhibits minor discrepancies in outcomes between the two datasets, with the CMDC dataset yielding results that are 0.4% higher in F1-scores. This marginal difference primarily stems from inherent variations in data composition between the two datasets: the CMDC dataset contains a more balanced distribution (56.5% healthy vs. 43.5% depressed), while the DAIC-WOZ dataset shows a greater imbalance (69% healthy vs. 31% depressed). Despite these distributional differences, the model maintains consistently strong performance across both datasets, demonstrating its robustness to variations in data characteristics.

t-SNE dimensionality reduction and visualization analysis

To further validate the influence of different modality feature on the classification effect, we visualized the extracted features using t-SNE dimensionality reduction. As shown in Fig. 3.

The visualization results indicate that unimodal data, when reduced to a 2D space, exhibits poor clustering performance, with text data performing particularly weakly. Upon further analysis, we found that this poor clustering in text data is largely due to the data augmentation process. During text slicing, certain segments lose their original semantic meaning, limiting the model’s ability to acquire sufficient information during training. This issue directly affects the recognition performance of the text modality, particularly in mental health applications, where textual data often underperforms relative to other modalities. For instance, short written expressions in depression-related datasets often fail to convey the emotional depth and nuance necessary for accurate emotion recognition. The evaluation results of the text modality, as presented in Tables 3 and 4, further corroborate this finding, highlighting the limitations of text-based features in emotion recognition tasks, especially when addressing complex mental health issues.

In contrast, the audio modality demonstrates superior clustering performance on the DAIC-WOZ dataset, showcasing a clear advantage in recognition accuracy over text. Audio data inherently carries richer emotional

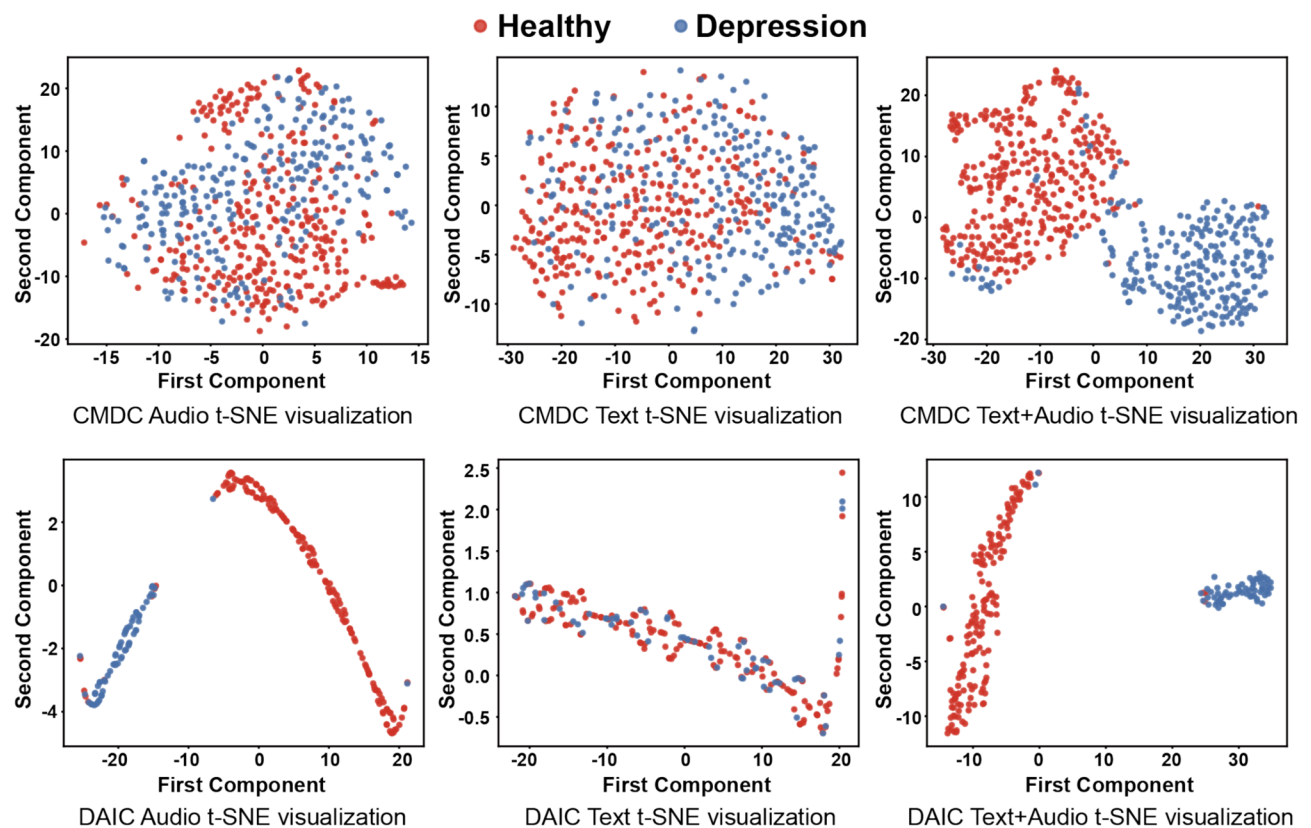


Fig. 3. t-SNE dimensionality reduction is applied, where red represents healthy subjects and blue represents depression patients in the figure.

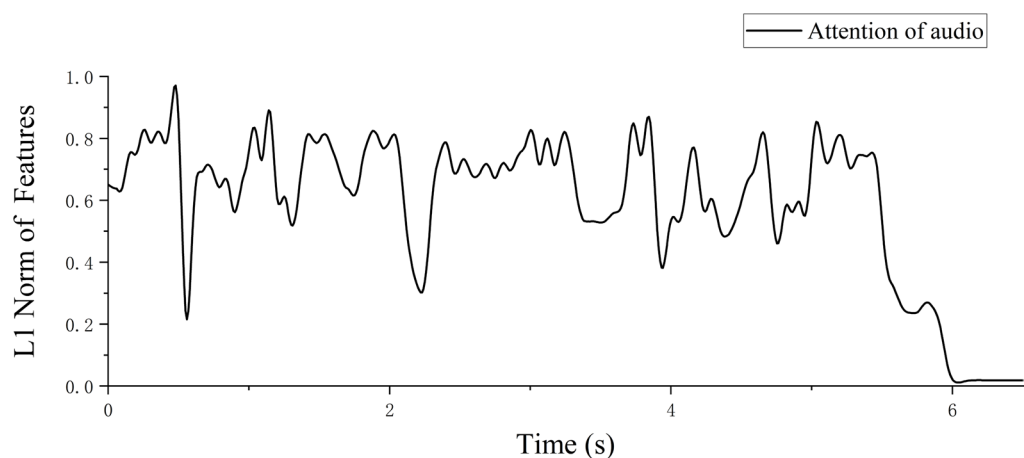


Fig. 4. The example of a single sample with 6.1 s of effective voice data length. Therefore, the subsequent data are no longer shown, and their characteristics are 0.

information, which significantly contributes to its better performance. The Wav2Vec 2.0 pre-trained model plays a key role in this success by effectively extracting features that capture subtle emotional expressions in audio, such as pitch, speaking rate, and pauses. These features allow the model to better recognize mental health-related emotions, including those associated with depression. This highlights the strength of Wav2Vec 2.0 in audio emotion analysis, as it excels at capturing nuanced emotional variations that are challenging to convey solely through text.

Despite the strengths of individual modalities, relying on a single modality is insufficient for comprehensively capturing complex emotions and psychological states. This underscores the importance of multimodal fusion approaches. On both the CMDC and DAIC-WOZ datasets, multimodal fusion significantly improves emotion

recognition accuracy, resulting in clearer clustering boundaries and enhanced discrimination capability. By employing MSC networks to expand the model's receptive field, features from different modalities are effectively integrated and processed. This integration enables the model to better capture emotional and mental health-related traits, particularly when addressing complex mental health issues.

The proposed framework enables detailed analysis of modality-specific feature contributions through quantitative evaluation of both speech and text representations. For the audio modality, we compute the L1 norm of temporal features to assess the relative importance of different speech segments. As visualized in Figure 4, this analysis reveals distinct patterns in feature weighting - brief pauses occurring at 0.68 seconds and 2.24 seconds demonstrate significantly lower weights compared to continuous speech segments. These findings align with clinical observations that sustained vocal characteristics (including prosodic variations, spectral changes, and phoneme transitions) contain more diagnostically relevant information than isolated pauses for depression assessment.

Text modality features are evaluated through their absolute magnitudes in the embedding space, with attention mechanisms further highlighting clinically significant linguistic markers. The analysis identifies specific lexical patterns that strongly influence model predictions (Fig. 5). Common function words and neutral expressions (e.g., positions 17, 42, and 44 in the sample text) contribute minimally, while self-reported symptoms and affect-related terms demonstrate substantially higher weights. Notably, phrases describing sleep disturbances (positions 7-10) emerge as particularly discriminative features, consistent with established diagnostic criteria for depressive disorders.

While the precise quantification of each modality's contribution remains challenging, our ablation studies (Tables 3 and 4) and feature visualizations (Fig. 3) consistently demonstrate that speech features exhibit greater discriminative power than text features for depression detection. This observation aligns with the clinical prominence of vocal biomarkers in mental health assessment, though the complementary value of textual information remains evident in the model's comprehensive evaluation framework.

In conclusion, unimodal approaches face inherent limitations in recognizing complex emotions and psychological states. The proposed multimodal fusion method overcomes these challenges by integrating the strengths of both audio and text, leading to significant improvements in recognition accuracy and classification performance. This approach demonstrates strong potential for applications in mental health, such as depression detection, where understanding nuanced emotional and psychological traits is critical.

Comparison with existing methods

For the CMDIC dataset, the classification results are compared with those of the attention-based multimodal fusion method⁶, as shown in Table 5. Through a comprehensive comparison of multiple evaluation criteria, the performance of our multimodal fusion method surpasses that reported in⁶. Our method demonstrates superior performance, particularly in terms of the F1 score, highlighting its effectiveness in depression classification.

Not limited to the CMDIC dataset, the proposed multimodal fusion method in this paper also demonstrates excellent performance on the DAIC-WOZ dataset, exhibiting high robustness and recognition accuracy. Compared with other multimodal depression detection methods that utilize the DAIC-WOZ dataset, the proposed method shows significant improvements across various metrics, including classification and regression. The comparison results are presented in Table 6.

Compared with other fusion methods, the proposed method demonstrates superior performance across both classification and regression tasks. As shown in Table 6, our approach outperforms existing methods by significant margins, with a 12.66% improvement in F1-score (96.66% vs 84.00%) compared to cross-attention baselines and an 11.66% enhancement (96.66% vs 85.00%) over LSTM-CNN architectures. These improvements stem from our optimized model structure that enables multi-scale information extraction while maintaining

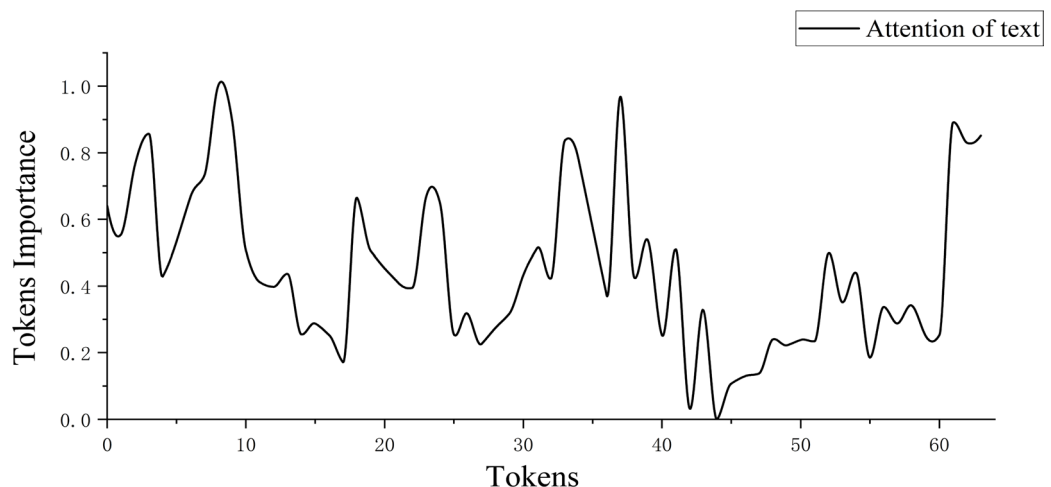


Fig. 5. The text data corresponds to the audio data, and the picture shows the attention characteristics of different tokens.

Metric	Ours	Cross-Attention ⁶
Accuracy	0.9747	–
Precision	0.9752	0.9668
Recall	0.9664	0.8605
F1 score	0.9708	0.9106
MAE	2.6232	3.6591
RMSE	4.0714	4.5872

Table 5. Comparison results on the CMDC dataset Significant values are in bold.

Metric	Ours	Cross-attention ⁶	LSTM ¹⁷	LSTM-CNN ¹⁸
Accuracy	0.9794	–	–	–
Precision	0.9702	0.8800	0.7100	0.7900
Recall	0.9631	0.8100	0.8300	0.9200
F1 score	0.9666	0.8400	0.7700	0.8500
MAE	2.8793	4.7400	5.1000	3.7500
RMSE	3.8199	5.8100	6.3700	5.4400

Table 6. Comparison results on the DAIC-WOZ Dataset Significant values are in bold.

computational efficiency. In the regression task for PHQ-8 score prediction, our method achieves the lowest error rates with MAE=2.8793 and RMSE=3.8199, representing reductions of 0.87 and 1.62 respectively compared to LSTM-CNN approaches.

For telehealth platforms, the improved accuracy enables more reliable remote screening, particularly valuable in resource-limited settings where in-person evaluations may be challenging. The model’s robust performance across different datasets suggests strong generalizability to diverse patient populations. The balanced performance across both classification and regression tasks demonstrates the framework’s versatility, making it suitable for deployment in various healthcare settings, from primary care offices to community mental health centers. These technical advantages position our method as a promising tool for enhancing depression screening workflows while maintaining clinical relevance and practical applicability. The combination of improved accuracy and robust performance across metrics addresses key requirements for implementation in real-world healthcare environments.

Conclusion

This study developed an efficient fusion model designed to capture temporal correlations and effectively utilize features from both voice and text. Departing from traditional feature engineering approaches, pre-trained models were integrated into an end-to-end architecture, enabling the direct input of preprocessed CMDC and DAIC-WOZ datasets for depression detection. This approach not only simplifies feature extraction for each modality but also expands the receptive field, significantly enhancing both the applicability and performance of the model. The evaluation results demonstrate that the proposed method consistently outperforms single-modal approaches in terms of recognition accuracy and PHQ-8 score prediction. These findings highlight the effectiveness and advantages of the proposed fusion model, offering a robust and reliable solution for automated depression detection.

However, the study has certain limitations. Due to dataset constraints, our work focused solely on the fusion of voice and text modalities. There are potential risks associated with data proportion issues, and the model may be prone to overfitting. Future research will explore the inclusion of additional modalities, such as video and image data, which could provide richer contextual and emotional information, thereby enhancing the model’s generalization and performance. Additionally, we will attempt to use larger datasets for training to mitigate the risk of overfitting. Another limitation lies in the computational complexity of self-supervised models such as BERT and Wav2Vec 2.0, which require significant memory resources, posing challenges for practical deployment. To address this, future research will investigate model compression techniques, such as model distillation and lightweight strategies, to reduce the number of parameters and memory usage, thereby improving the model’s efficiency and suitability for real-world applications. In conclusion, this study demonstrates the potential of multimodal fusion approaches in advancing automated depression detection. By effectively integrating voice and text features and leveraging state-of-the-art pre-trained models, the proposed method provides a robust foundation for future research in mental health diagnostics and related applications.

Data availability

The data of CMDC that support the findings of this study are available in the School of Computer and Communication Engineering, University of Science and Technology Beijing, [https://iee-dataport.org/open-access/chinese-multimodal-depression-corpus]. In addition, the data of DAIC-WOZ are available in the Institute for Creative Technologies, University of Southern California, [https://dcapwoz.ict.usc.edu/]. These datasets are

publicly available datasets, but permission is required before use. If you want to get the data, ask for permission at the two addresses above.

Code availability

The source code for all the experiments can be viewed in <https://github.com/xzrlzz/A-multimodal-model-for-detecting-depression-using-text-and-audio-data>.

Received: 6 February 2025; Accepted: 21 May 2025

Published online: 01 July 2025

References

- Huang, Y. et al. Prevalence of mental disorders in China: A cross-sectional epidemiological study. *Lancet Psychiatry* **6**, 211–224. [https://doi.org/10.1016/S2215-0366\(18\)30511-X](https://doi.org/10.1016/S2215-0366(18)30511-X) (2019).
- Eid, M. M., Yundong, W., Mensah, G. B. & Pudasaini, P. Treating psychological depression utilising artificial intelligence: AI for precision medicine-focus on procedures. *Mesopotamian J. Artif. Intel. Healthc.* **2023**, 76–81. <https://doi.org/10.58496/MJAIH/2023/015> (2023).
- Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Pike, C. T. & Kessler, R. C. The economic burden of adults with major depressive disorder in the united states (2005 and 2010). *J. Clin. Psychiatry* **76**, 5356. <https://doi.org/10.4088/JCP.14m09298> (2015).
- Stecher, C., Cloonan, S. & Domino, M. E. The economics of treatment for depression. *Ann. Rev. Public Health.* <https://doi.org/10.1146/annurev-publhealth-061022-040533>. (2023).
- Sun, H. et al. Multi-modal adaptive fusion transformer network for the estimation of depression level. *Sensors* **21**, 4764. <https://doi.org/10.3390/s21144764> (2021).
- Zou, B. et al. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders[dataset]. *IEEE Transactions on Affective Computing* **14**, 2823–2838. <https://doi.org/10.1109/TAFFC.2022.3181210> (2022).
- Squires, M. et al. Deep learning and machine learning in psychiatry: A survey of current progress in depression detection, diagnosis and treatment. *Brain Informatics* **10**, 10 (2023).
- Manjulatha, B. & Pabboju, S. Multimodal depression detection using deep learning in the workplace. in *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 1–8 (IEEE, 2024). <https://doi.org/10.1109/ICAECT60202.2024.10468966>.
- Cinaz, B., Arnrich, B., La Marca, R. & Tröster, G. Monitoring of mental workload levels during an everyday life office-work scenario. *Pers. Ubiquitous Comput.* **17**, 229–239. <https://doi.org/10.1007/s00779-011-0466-1> (2013).
- Bylsma, L. M., Morris, B. H. & Rottenberg, J. A meta-analysis of emotional reactivity in major depressive disorder. *Clin. Psychol. Rev.* **28**, 676–691. <https://doi.org/10.1016/j.cpr.2007.10.001> (2008).
- Krause, F. C., Linardatos, E., Fresco, D. M. & Moore, M. T. Facial emotion recognition in major depressive disorder: A meta-analytic review. *J. Affect. Disord.* **293**, 320–328. <https://doi.org/10.1016/j.jad.2021.06.053> (2021).
- Amanat, A. et al. Deep learning for depression detection from textual data. *Electronics* **11**, 676. <https://doi.org/10.3390/electronics11050676> (2022).
- Islam, M. R. et al. Depression detection from social network data using machine learning techniques. *Health Inf. Sci. Syst.* **6**, 1–12. <https://doi.org/10.1007/s13755-018-0046-0> (2018).
- Orabi, A. H., Buddhitha, P., Orabi, M. H. & Inkpén, D. Deep learning for depression detection of twitter users. in *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, 88–97 (2018). <https://doi.org/10.18653/v1/W18-0609>.
- Mamidiseti, S. & Reddy, A. M. A stacking-based ensemble framework for automatic depression detection using audio signals. *Int. J. Adv. Comput. Sci. Appl.* (2023).
- Zhang, P., Wu, M., Dinkel, H. & Yu, K. Depa: Self-supervised audio embedding for depression detection. in *Proceedings of the 29th ACM international conference on multimedia*, 135–143 (2021). <https://doi.org/10.1145/3474085.3479236>.
- Al Hanai, T., Ghassemi, M. M. & Glass, J. R. Detecting depression with audio/text sequence modeling of interviews. in *Interspeech*, 1716–1720 (2018).
- Lin, L., Chen, X., Shen, Y. & Zhang, L. Towards automatic depression detection: A bilstm/1d cnn-based model. *Appl. Sci.* **10**, 8701. <https://doi.org/10.3390/app10238701> (2020).
- Zhang, Z. et al. Multimodal sensing for depression risk detection: Integrating audio, video, and text data. *Sensors* **24**, 3714. <https://doi.org/10.3390/s24123714> (2024).
- Ye, J. et al. Multi-modal depression detection based on emotional audio and evaluation text. *J. Affect. Disord.* **295**, 904–913. <https://doi.org/10.1016/j.jad.2021.08.090> (2021).
- Huang, X. et al. Depression recognition using voice-based pre-training model. *Sci. Rep.* **14**, 12734. <https://doi.org/10.1038/s41598-024-63556-0> (2024).
- Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **33**, 12449–12460 (2020).
- Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018).
- Yang, G. et al. Afpn: Asymptotic feature pyramid network for object detection. in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2184–2189 (IEEE, 2023).
- Han, X., Chen, F. & Ban, J. Fmfn: A fuzzy multimodal fusion network for emotion recognition in ensemble conducting. *IEEE Trans. Fuzzy Syst.* (2024).
- Sun, P., Zhang, Y., Liu, Z., Chen, D. & Zhang, H. More than vanilla fusion: a simple, decoupling-free, attention module for multimodal fusion based on signal theory. arXiv preprint [arXiv:2312.07212](https://arxiv.org/abs/2312.07212) (2023).
- Tsai, Y.-H. H. et al. Multimodal transformer for unaligned multimodal language sequences. in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019, 6558 (2019).
- DeVault, D. et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support[dataset]. in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 1061–1068 (2014).

Acknowledgements

The authors would like to thank Zhengzheng Lin and Yilong Liao for their generous assistance during the preparation of the paper.

Author contributions

All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by Z.X., Y.G., F.W., L.Z., J.W., and J.S. The first draft of the manuscript was written by Z.X. and L.Z. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. Z.X. and Y.G. are co-first authors of the paper. L.Z. is the corresponding author.

The datasets used in this study were approved by the respective authorities overseeing the CMDC and DAIC-WOZ datasets. All methods were carried out in accordance with the ethical guidelines and regulations of these authorities. Informed consent was obtained from all study participants. Additionally, other experiments involving these two datasets were approved by the Research Ethics and Science and Technology Safety Committee of South-Central Minzu University (Approval No. 2024-scec-106).

Funding

This work was supported by the Fundamental Research Funds for the Central Universities of South-Central Minzu University [Grant Number CZQ23031].

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025