WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 5, Number 1, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v5n1.16
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Multimodal Data-Based Text Generation Depression Classification Model

Shukui Ma [1], Pengyuan Ma [1], Shuaichao Feng [1], Fei Ma [2], Guangping Zhuo [1, *]

[1] School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China

[2] Department of Computer Science and Technology, Taiyuan University, Taiyuan 030032, China

ABSTRACT

Depression classification often relies on multimodal features, but existing models struggle to capture the similarity between multimodal features. Moreover, the social stigma surrounding depression leads to limited availability of datasets, which constrains model accuracy. This study aims to improve multimodal depression recognition methods by proposing a Multimodal Generation-Text Depression Classification Model. The model introduces a Multimodal-Deep-Extract-Feature Net to capture both long- and short-term sequential features. A Dual Text Contrastive Learning Module is employed to generate emotionally salient word embeddings from patients' transcribed text. Contrastive learning brings similar features closer and pushes dissimilar features apart, thereby enhancing the representation of dual-text features. Finally, a Joint Multi-modal Fusion Attention mechanism is proposed to amplify the representation of dominant modalities, effectively integrate all modalities, and capture global multimodal features. This integrated approach improves depression recognition accuracy, facilitating timely intervention and support for patients. The model achieves accuracy rates of 89.5% on the DAIC-Woz dataset and 92% on the MDD2024 dataset.

KEYWORDS

Depression Classification; Multimodal; Dual Text Contrastive Learning Module; Joint Multi-modal Fusion Attention

1. INTRODUCTION

According to statistics from the World Health Organization (WHO), it is estimated that 3.8% of the global population suffers from depression, including 5% of adults (4% of males and 6% of females), as well as 5.7% of adults aged 60 and above. It is estimated that approximately 280 million people worldwide are affected by depression. The incidence rate of depression in women is approximately 50% higher than that in men. Globally, over 10% of pregnant women and women who have just given birth are affected by depression [1]. Each year, more than 700,000 people die by suicide, making it the fourth leading cause of death among individuals aged 15-29. According to a recent study published in The Lancet, depression has significantly increased in 2020 due to the impact of the COVID-19 pandemic [2]. The WHO estimates that by 2030, depression will become the second leading cause of disease burden [3]. However, the treatment of Major Depressive Disorder (MDD) is costly, and there is a shortage of trained professionals, with over 75% of patients in low- and middle-income countries not receiving treatment [4].

In response to this demand, recent progress has been made in the identification of depression using deep learning techniques, both for processing unimodal [5-8] and multimodal data [9-15]. For example, the use of short-term speech segments [16] or the integration of various acoustic features

through deep learning models [17] has shown promising results. Furthermore, Temporal Convolutional Networks have been employed to analyze multimodal data, such as facial expressions, speech signals, and transcribed text, thereby improving diagnostic accuracy [18].

Despite these advancements, integrating data from interactions between patients and doctors remains challenging due to the dynamic and inconsistent nature of different modalities, particularly in capturing subtle expression features. Additionally, the scarcity of Chinese depression datasets, given the sensitivity and privacy of depression, contributes to low model accuracy and the risk of overfitting.

In order to address the aforementioned issues, this study proposes a novel network architecture called Multimodal-Deep-Extract-Feature Net (MDEFNet). Building upon the existing TCN architecture [19], this network incorporates the gated mechanism of LSTM [20]. By introducing gate functions in the unit structure, this model not only effectively captures the features between dialogue paragraphs of patients and doctors but also addresses the challenge of long-term dependencies. To tackle the problem of limited datasets and potential overfitting, a dual text contrastive learning module (DTCL) is introduced in this study. The core idea of this module is to utilize a generative text model to perform synonymous replacements on words expressing patients' psychological activities in the original text. Through contrastive learning, the generated text is encouraged to have similar features in the word vector space as the original text, while different features are dispersed. Additionally, by adjusting standardization and temperature parameters, we further enhance the representation effectiveness of text features. To effectively capture the weak feature expressions of each modality, this study introduces a Joint Multi-modal Fusion Attention mechanism (JMFATT). This attention mechanism amplifies the feature expressions of weaker modalities by leveraging the stronger modalities, enabling the effective capturing of subtle features. Moreover, the Joint Multi-modal Fusion Attention mechanism can effectively fuse all modality features to capture global features in multi-modal data.

After multiple experimental analyses, the proposed model achieved an accuracy of 89.5% and a recall rate of 89.5% on the publicly available DAIC-WOZ dataset. Furthermore, to address the scarcity of Chinese depression databases, this study collaborated with various hospitals, obtained patient consent, and constructed the MDD2024 dataset. To protect patient privacy, this dataset does not include any information related to patient privacy.

In summary, the contributions of this study are as follows:

(1) This study introduces a feature extraction network framework called Multi-Modal Deep Extract Feature Net (MDEFNet). Building upon TCN, this network incorporates LSTM gated units to effectively capture global features of long-term dialogues between patients and doctors without losing granular-level features.

(2) This study innovatively introduces a dual text contrastive learning module in the network, named Dual Text Contrastive Learning (DTCL). The module is capable of taking in dual text features simultaneously and performing contrastive learning through a contrastive loss function, which amplifies feature representation by bringing similar features closer in the word vector space.

(3) To address the issue of strong and weak feature representation in multi-modal experiments, a novel attention mechanism named Joint Multi-modal Fusion Attention (JMFATT) is proposed in this study. This mechanism leverages the strong feature representation of one modality to enhance the feature representation of another modality that is comparatively weaker. Additionally, it facilitates the fusion of different modalities and captures the similarities between them during the fusion process, thereby improving the accuracy of the model.

(4) This study actively collaborates with local hospitals to construct the publicly available Chinese depression dataset MDD2024, while ensuring the protection of patient privacy. This initiative aims to contribute significantly to the advancement of Chinese depression recognition.

2. MODEL

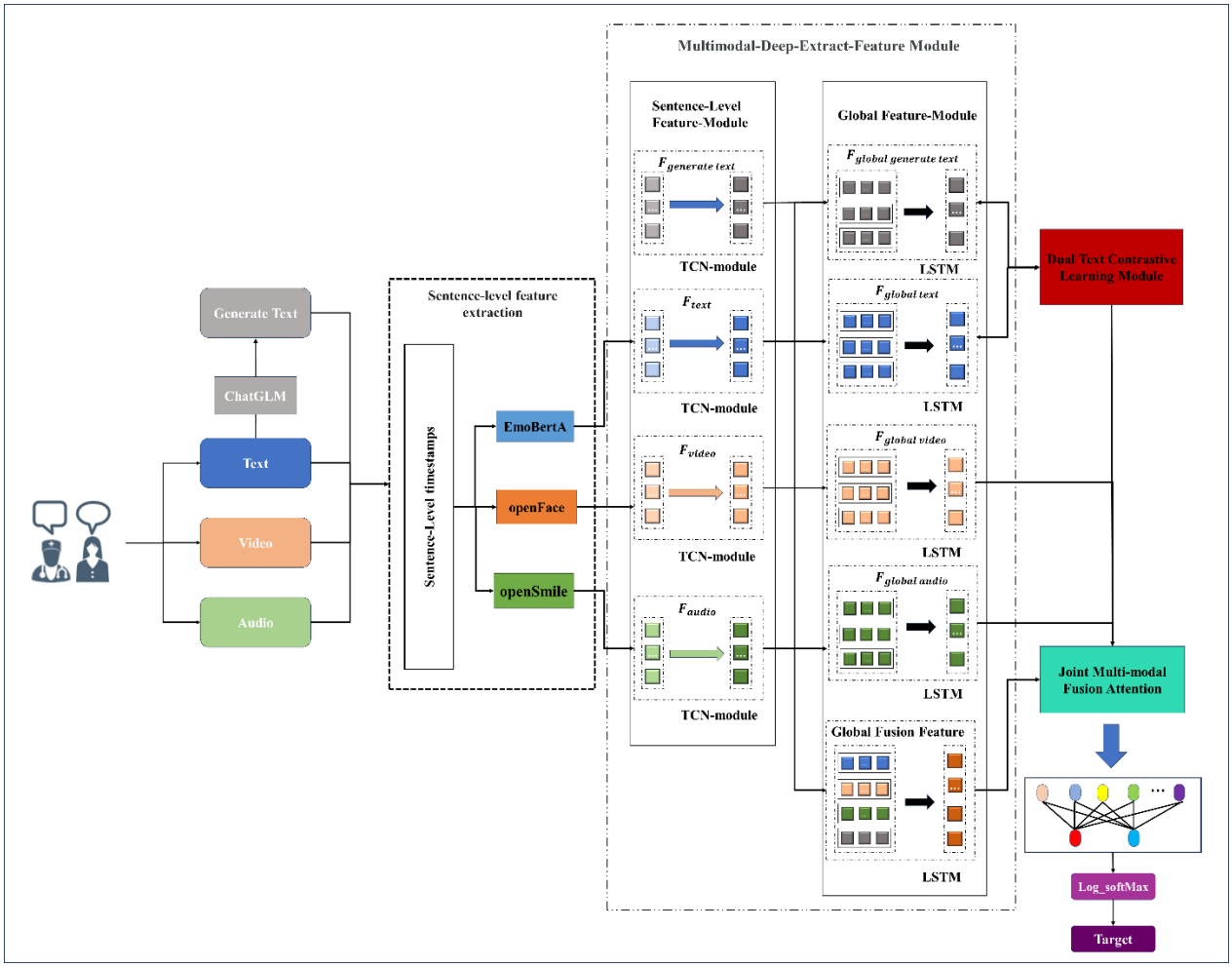

Figure 1. Multimodal Generation- Text Depression Classification Model

In order to address the challenges in multi-modal depression classification, this paper proposes a novel depression classification network called Multi-Modal Generation Text Depression Classification Model (MGTDCM), as shown in Figure 1. To tackle the issue of capturing global features during the extraction of patient data, the network introduces MDEFNet (Multi-Modal Deep Extract Feature Net). MDEFNet utilizes four modal features: temporal facial features, patient audio features, interview transcript text features, and generated text features. These modal features are aligned based on timestamps and segmented at the sentence-granularity level. The segmented modules are then fed into MDEFNet for deep feature extraction. Firstly, the interview transcript text is replaced with generated text features that reflect the patient's psychological activities using the ChatGLM model [21]. Next, the sentence-granularity features are extracted using TCN, adjusting the number of convolutional layers and dilation factors to flexibly adjust the size of the receptive field and retain feature representation. The processed features are concatenated into global features and processed by an LSTM model to effectively utilize contextual information and understand the semantic relationships between sentences. During the diagnosis process, to gain a deeper understanding of the patient, thresholds are introduced and a limit is set on the number of dialogues to reduce model complexity while ensuring the timeliness of the data. To address overfitting issues caused by limited datasets, this paper adopts the Dual Text Contrastive Learning (DTCL) method to enhance the expressive power of text features. By using the deep-extracted dual text features as the data source and employing contrastive learning loss functions to calculate the similarity between the dual texts, similar features are brought closer together in the embedding space while dissimilar features are pushed further apart, amplifying the expression of similar features in the dual texts.

Finally, the Joint Multi-modal Fusion Attention (JMFATT) is utilized to address the issue of varying feature expression in different modalities. JMFATT leverages the strong feature representation of one modality to amplify the feature expression of another modality that has comparatively weaker representation. It also facilitates the fusion of different modalities, capturing the similarities between them, thereby enhancing the accuracy of the model.

2.1. Multi-Modal Deep Extraction Feature Network

In depression detection, the analysis of temporal data is crucial as depression exhibits significant time dependence. Time Convolutional Networks (TCN) and Long Short-Term Memory Networks (LSTM) are two commonly used models for processing temporal data, each having unique advantages. TCN excels in capturing short-term dependencies, with its convolutional structure being able to quickly process input data and maintain the order of information through causal convolutions. Its parallel processing capabilities make it more efficient when dealing with longer sequences, effectively extracting local features and capturing instantaneous emotional changes. On the other hand, LSTM networks have a significant advantage in modeling long-term dependencies. Through the mechanisms of forget gates, input gates, and output gates, it effectively preserves important information while suppressing irrelevant information, allowing for a better understanding and prediction of long-term mood changes. This is crucial in depression detection as a patient's emotional state is often influenced by their past experiences. However, using TCN or LSTM alone has limitations in processing complex temporal data. Thus, this study proposes MDEFNet that integrates the strengths of TCN and LSTM to improve the ability to extract and understand complex temporal features. The structure of MDEFNet is shown in Figure 2:

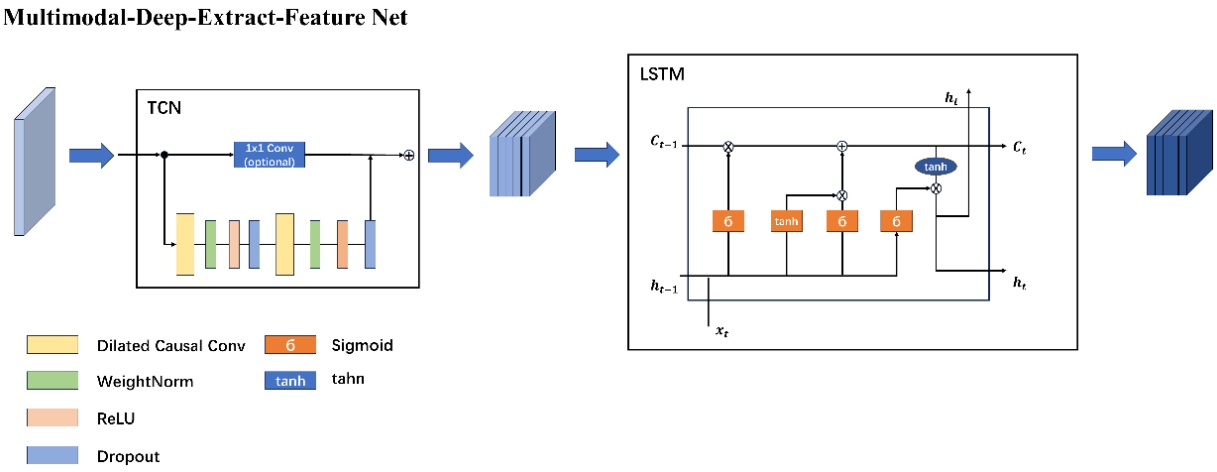

Figure 2. Multi-Modal Deep Extraction Feature Network Architecture

In MDEFNet, the sentence-granularity features are first fed into the Time Convolutional Network (TCN) module for deep feature extraction. The TCN model adopts structures such as causal convolutions, dilated convolutions, and residual blocks to control the receptive field size and limit the number of parameters, ensuring an acceptable model complexity. Among these structures, causal convolution is a key component of TCN, and its formula is as follows:

$$y[t] = \sum_{k=0}^{t} x[t-k] \bullet \omega[k] \tag{1}$$

$y[t]$ represents the output at time and $\sum_{k=0}^{k-1} x[t-k]$ is the sum of inputs before time. $\omega[k]$ indicates the weight of the convolutional kernel. At t time, the output $y[t]$ only depends on the current input

$x[t-k]$ and the inputs x from the past, while it is independent of future inputs. This avoids the influence of subsequent inputs on the current output. Furthermore, the receptive field of causal convolution is limited. When there are fewer convolutional layers, it can only capture information from a shorter historical range for predicting the current time step. To address this issue, the TCN model introduces dilated convolutions to expand the receptive field. The formula for dilated convolution is as follows:

$$y[t] = \left(X *_d f\right) = \sum_{i=0}^{n-1} f(i) \cdot X_{T-d \cdot i} \tag{2}$$

In this context $y[t]$ denotes the output signal, while X denotes the output signal, and d is the dilation factor. ${}_d f$ denotes the convolution kernel transformed by the dilation factor d, $x_{T-d \cdot i}$ corresponds to the past samples of the input signal associated with the current time step T, Each input sample x is dilated and combined with the convolution kernel $f(i)$ through a weighted summation to achieve the effect of dilated convolution. However, very deep networks may suffer from issues such as training instability and gradient vanishing. To address these problems, the TCN model adopts a residual block structure, and its formula is as follows:

$$X^i = \delta\left(F\left(X^{(i-1)}\right) + X^{(i-1)}\right) \tag{3}$$

The i-th residual block is composed of the output of the previous residual block $X^{(i-1)}$ and the original input. One branch performs a dilated convolution operation, while the other branch uses a 1×1 convolution kernel to upsample or downsample the original input data, ensuring that the dimensions of the convolved data can be directly added to produce the output $X^{(i-1)}$. After completing the TCN processing, sentence-level features are concatenated into global features and passed as input to the Long Short-Term Memory (LSTM) model. When processing long-duration data, the LSTM module offers several advantages over the TCN. First, through its complex memory units and gating mechanisms, LSTM can flexibly capture and retain long-term dependencies in data sequences. Among them, the LSTM contains three primary gate structures, with the forget gate determining the extent to which the hidden state from the previous time step is retained. The formula is as follows:

$$f_t = \sigma(\omega_f \bullet [h_{t-1}, x_t] + b_f) \tag{4}$$

In this context, f_t represents the output of the forget gate, σ denotes the sigmoid activation function, and the hidden state of the previous layer is represented by h_{t-1}. x_t represents the input features, and ω_f is the weight matrix. In the forget gate, h_{t-1} and x_t are combined through a weighted summation using the weight matrix, ultimately yielding f_t. When the value of f_t approaches 0, it indicates that more information should be forgotten, which helps capture short-term dependencies in the data. The input gate determines which parts of the current input information will be added to the memory cell. The formula is as follows:

$$i_t = \sigma(\omega_i \bullet [h_{t-1}, x_t] + b_i) \tag{5}$$

The variable i_t represents the output of the input gate. When its value approaches 1, it indicates that more data needs to be retained. Conversely, when it approaches 0, it signifies that more data needs to be forgotten. This mechanism aids in capturing long-term dependencies within the data. The output gate determines how much information is output from the memory cell. The formula is as follows:

$$o_t = \sigma(\omega_o \bullet [h_{t-1}, x_t] + b_o) \tag{6}$$

In this context, o_t represents the output of the output gate. A value closer to 1 indicates that more information is being transmitted to the subsequent layer, while a value closer to 0 signifies that less information is being conveyed. In addition to the fundamental structure comprising three gates, the model incorporates two primary memory units: the candidate memory unit and the update memory unit, whose formulas are presented in Equation 7 and Equation 8, respectively.

$$C_t = \tanh(\omega_c \bullet [h_{t-1}, x_t] + b_c) \tag{7}$$

$$C_t = f_t \bullet C_{t-1} + \mathrm{i}_t \bullet C_t \tag{8}$$

The output of the candidate memory cell is denoted as C_t, and the current memory cell state is represented as C_t. The candidate memory cell state provides the potential value of new information, while the memory cell state preserves filtered information. The combination of these two mechanisms enables LSTM to effectively manage the flow of information in time series data, enhancing its ability to capture both long-term and short-term dependencies.

By combining TCN and LSTM in a synergistic manner, MDEFNet fully leverages the advantages of these two types of temporal networks, including efficient time feature extraction and comprehensive management of model complexity, while effectively capturing long-term dependencies and mitigating gradient-related challenges. This dual-network architecture makes the model a powerful tool for improving the representation of patient data features, thereby enhancing the diagnostic capabilities of depression classification.

2.2. Dual Text Contrastive Learning

To address the issue of overfitting caused by the scarcity of depression dataset, this study introduces a new modality of generating text as input to the model, building upon the current mainstream research. This modality specifically focuses on the sentences in patients' transcription texts that reflect their psychological activities and performs synonym replacement. To fully utilize this modality, this study incorporates a Dual Text Contrastive Learning module, as shown in Figure 3:

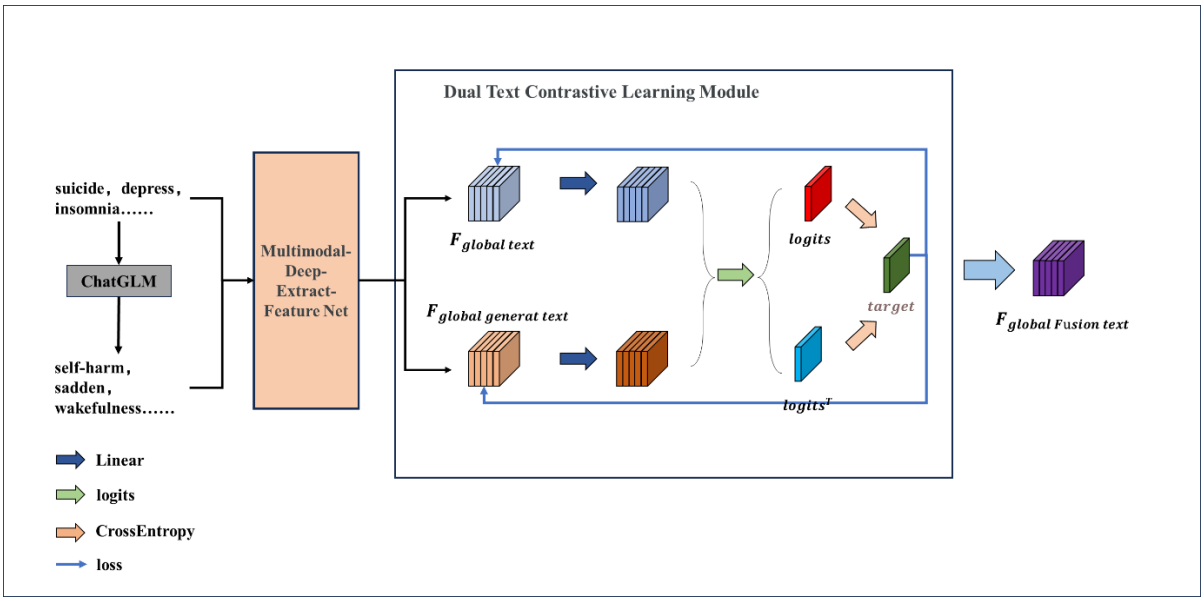

Figure 3. Dual Text Contrastive Learning Module

The Dual Text Contrastive Learning module adopts a contrastive learning approach to enhance the feature representation capability of both the text and its generated transcription. Specifically, by using a corpus containing multiple texts, the generated transcription is compared to the original text through a contrastive loss function to enhance the feature representation learning between them. The goal of this loss function is to maximize the similarity between positive pairs while minimizing the similarity between negative pairs, thereby improving the accuracy of the model. The steps are as follows:

Feature normalization is performed by using L2 normalization on the feature vectors, ensuring that the length of each feature vector is 1. This normalization guarantees that the feature vectors have a unit length, eliminating the impact of absolute feature values on similarity calculations. The formula is as follows:

$$Feature_i^{normalized} = \frac{Feature_i}{\left\| Feature_i \right\|} \tag{9}$$

Logits calculation: The dot product computes the cosine similarity between two feature vectors and is multiplied by a temperature parameter to adjust the distribution of logits. The formula is as follows:

$$\mathrm{logits} = Feature_1^{normalized} \bullet (Feature_2^{normalized})^{\mathrm{T}} \bullet e^{t} \tag{10}$$

Here, T represents the temperature parameter, which controls the concentration of the similarity scores. Higher temperature values make the similarity distribution smoother, while lower temperature values make the distribution more concentrated, thereby affecting the learning effectiveness of the model.

Contrastive loss function, the core idea of the contrastive loss function is to compute the cross-entropy loss to obtain the contrastive loss. This process aims to encourage the model to learn better feature representations. The formula is as follows:

$$L_{con} = \frac{CrossEntropy(\mathrm{logits}, \mathrm{labels}) + CrossEntropy(\mathrm{logits}^{\mathrm{T}}, \mathrm{labels})}{2} \tag{11}$$

In the numerator, the cross-entropy loss is calculated separately for the positive and negative pairs, where the labels are used to indicate the correct pairings. For each pair of features, the labels are the same for positive pairs and different for negative pairs. The final loss is the average of these two losses, aiming to ensure that the model learns equally from positive and negative pairs.

2.3. Joint Multi-modal Fusion Attention

In multi-modal experiments, different modalities may have different feature representations, with some modalities expressing stronger features and others exhibiting relatively weaker features. Moreover, in the process of multi-modal fusion, it is often difficult to accurately capture the similarities between different modalities. To address this problem, we propose a Joint Multi-modal Fusion Attention (JMFATT) mechanism, whose structure is shown in Figure 4:

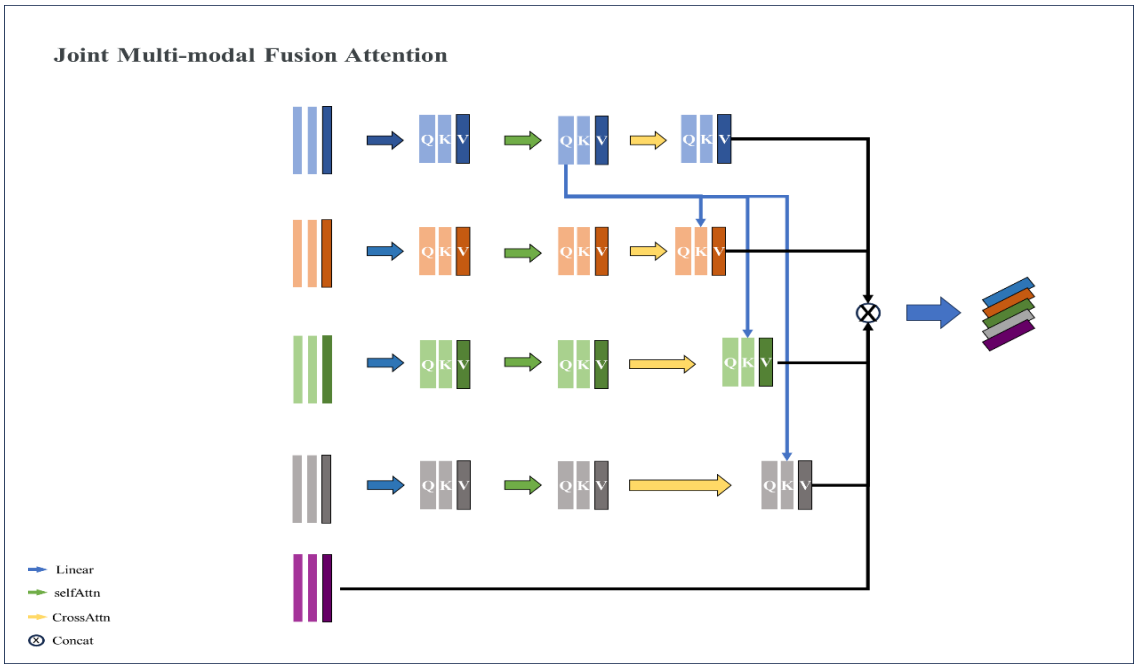

Figure 4. Joint Multi-modal Fusion Attention

The core idea of JMFATT is to use the modalities with strong feature representations to influence the modalities with weaker feature representations, in order to capture subtle features that are difficult to capture otherwise. This attention mechanism allows for the effective extraction of similar features among different modalities, thereby improving the accuracy of the model. Finally, by using this attention mechanism, all modalities are fused to capture the similarity features between different modalities. This approach effectively enhances the accuracy of the model in multi-modal depression classification. The steps are as follows:

Feature projection: Each input feature is linearly transformed to ensure that subsequent attention calculations can be performed in the same space. The formula is as follows:

$$Q_i, K_i, V_i = \mathrm{Linear}(F_{\mathrm{input}}) \tag{12}$$

Where F_{input} represents the input feature, Q_i represents the query, K_i represents the key, and V_i represents the value.

Self-attention calculation: To capture the internal correlations within each modality, the self-attention mechanism is used to compute the self-attention for each modality. It generates attention scores by taking the dot product between the query and the key, applies the softmax function to obtain the weights, and then computes the weighted sum of the values. This allows the model to identify important features within each modality. The formula is as follows:

$$\mathrm{SelfAttention}(Q_i, K_i, V_i) = \mathrm{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{13}$$

Here, d_k represents the dimensionality of the key vectors. Specifically, d_k is used to scale the dot product between the query and the key in order to avoid large values that may affect the stability of the softmax function when computing attention scores.

Cross-modal attention: To capture the relationships between different modalities, a query from one modality is used to attend to the values from other modalities. Specifically, a strong modality is

selected as the query to focus and amplify the feature representations of weaker modalities. Meanwhile, the model uses cross-modal attention to better support the strong modality with information from other modalities. This mechanism allows the model to flexibly adjust the interaction between modalities in different tasks, enhancing weaker features and amplifying stronger features, ultimately improving overall performance. The formula is as follows:

$$\text{CrossAttention}(Q_i, K_i, V_i) = \text{Softmax}(\frac{Q_i K_j^T}{\sqrt{d_k}})V_j \tag{14}$$

In this context, Q_i represents the query from the strong modality, while K_j and V_j represent the key and value from the weaker modality, respectively.

Output fusion and projection: The outputs from self-attention and cross-modal attention are combined and mapped to the final output space. The self-attention outputs for each modality are weighted and combined with the cross-modal attention outputs from other modalities. Finally, a linear function is applied to project the fused representation. The formula is as follows:

$$\text{Output} = \sum_i \text{Linear}(\gamma_i \text{SelfAttn}_i + \sum_{i \neq j} \delta_{ij} \bullet \text{CrossAttn}_{i,j}) \tag{15}$$

In this context, γ_i and δ_{ij} are dynamic fusion weights that are adjusted based on the model's classification results to dynamically tune the contributions of each modality. The fusion weights determine how much importance is given to the self-attention output and the cross-modal attention output for each modality.

2.4. Loss Function

The loss function plays a crucial role in training machine learning models. It not only evaluates the model's performance but also guides the learning process and parameter adjustments. In this framework, the cross-entropy loss function is initially employed, which is commonly used for classification tasks. However, due to the class imbalance in the dataset, weights are introduced to make the model pay more attention to the minority class. The formula for the weighted cross-entropy loss function is as follows:

$$\text{L}_{\text{cross}} = -\frac{1}{N} \sum_{i=1}^{N} [\omega_i \bullet \{y_i \log(p_i) + (1 - y) \log(1 - p_i)\}] \tag{16}$$

Here, N represents the number of samples, y_i represents the true labels, p_i represents the model's predicted probabilities, and ω_i represents the weight for class (i), which is used to adjust the impact of different classes on the loss. Next, the loss function of the contrastive learning, denoted as L_{con}, is added to the equation:

$$\text{loss} = \text{L}_{\text{cross}} + \text{L}_{\text{con}} \tag{17}$$

Indeed, this combination of binary cross-entropy loss and contrastive learning loss achieves joint optimization and comprehensive learning, which helps improve the quality of feature representation and enhance the overall performance of the model.

3. EXPERIMENTAL DATA

3.1. Dataset

In this study, we mainly used the publicly available Chinese depression dataset MDD2024, which was constructed in collaboration with local hospitals in Taiyuan, China. We also conducted validation experiments using the publicly available depression dataset DAIC-WOZ. The MDD2024 dataset consists of digital conversations between doctors and patients, where doctors asked sensitive questions and simulated conversations to facilitate a more realistic interview scenario. The dataset includes data from 100 participants who voluntarily participated in the study with informed consent. No data that could potentially compromise patient privacy is included in the dataset. The publicly available DAIC-WOZ dataset includes audio, video, and transcript features of both non-depressed and depressed individuals. The Patient Health Questionnaire-8 (PHQ-8) [22] was used to assess depression in patients. Through analysis of the MDD2024 dataset, we found that the gender distribution of patients is shown in Table 1 below.

Table 1. Gender and Age Distribution Statistics of Depressed Patients and Non-Depressed Individuals

Group	Gender	Elderly	Middle aged	Young Adults	Minors	Total
Depressed	Male	0	5	11	1	17
	Female	3	18	28	1	50
Non-depressed	Male	1	2	4	1	8
	Female	0	7	13	5	25
Total		4	32	56	8	100

The age groups were classified according to international standards, with individuals aged 60 and above classified as elderly, those aged 40 to 60 classified as middle-aged, those aged 18 to 39 classified as young adults, and those under 18 classified as minors. As shown in Table 1, young adults accounted for 56% of the participants in our study, while only 4% of patients were elderly. The proportions of the other age groups were 32% and 8%, respectively. Depression was evaluated using the PHQ-8 assessment scale, with a score of 9 or higher considered as the standard for depression. To minimize differences in language ability, only participants whose native language was Mandarin were included in the study. Due to the limited number of MDD (Major Depressive Disorder) patients, there were more participants in the control group, and their age range was wider. If matching of the control group is required, a subset of the dataset can be selected to minimize the influence of gender and increase statistical power [11].

3.2. Data Preprocessing

This study utilized two main datasets, namely the self-collected MDD2024 dataset and the publicly available DAIC-WOZ dataset, to facilitate robust analysis of depression indicators. The MDD2024 dataset and the DAIC-WOZ dataset contain 100 and 192 cases, respectively, with 67 and 74 confirmed positive cases of depression. To ensure the methodological rigor of model training and validation, the two datasets were split into a 8:2 training and testing ratio. This structured split ensures a balanced representation of depression cases in both datasets, enhancing the reliability and effectiveness of the research results. In particular, for the MDD2024 dataset, 80 cases were allocated to the training set, including 54 depression patients, while 20 cases were allocated to the testing set, including 13 depression cases. To address the significant imbalance between depression and non-depression cases, weighted ratios were implemented in the study. These ratios were calculated based on the differences in case numbers to ensure fair representation and enhance the robustness of the analysis of research results. Specifically, the weight calculation formula is as follows:

$$weight = [\frac{S_t}{N_0}, \frac{S_t}{N_1}] \tag{19}$$

Where S_t represents the total number of samples in the training set, N_0 represents the number of negative samples in the training set, and N_1 represents the number of positive samples in the training set. To protect patient privacy, this study implemented a series of measures to safeguard patient confidentiality. Firstly, personal sensitive information such as names, dates, and locations has been removed from recorded audio and transcribed text. Secondly, in the case of video data, only 3D facial landmarks containing 68 pixels have been retained, which are crucial for measuring facial movements such as eye, mouth, and head movements. This ensures that there is not enough information available for personal identification, but sufficient information for measuring facial movements.

3.3. Feature Extraction for Each Modal

3.3.1. Facial Feature Extraction

The steps typically involved in image feature extraction are as follows: Firstly, the video data is segmented into different segments based on the timestamps of the audio data, and unnecessary data is removed to improve the efficiency of the model's operation. Next, the OpenFace model [23] [23] is used to extract keypoint features from the patient's face, covering 68 key points including lip movements, facial expressions, and eye movements. The keypoint extraction is performed with a step size of 0.033 seconds. These keypoint features provide valuable information about facial movements and expressions, which can be further utilized for analysis and interpretation in the context of the study.

Figure 5. Facial features were extracted using the OpenFace tool

3.3.2. Audio Feature Extraction

The audio data consists of complete interview audio for each patient. Initially, the original interview audio is segmented into smaller units based on the timestamp of each sentence spoken by the patient, forming audio data at the sentence level. Then, the segmented audio units are passed through the openSmile encoder [24] to extract features based on the patient's audio intonation variations, sound intensity, fundamental frequency, spectral characteristics, and Mel-frequency cepstral coefficients (MFCCs). By extracting these features, valuable information about the patient's audio characteristics, such as pitch, loudness, and spectral properties, can be captured and utilized for further analysis and interpretation in the study.

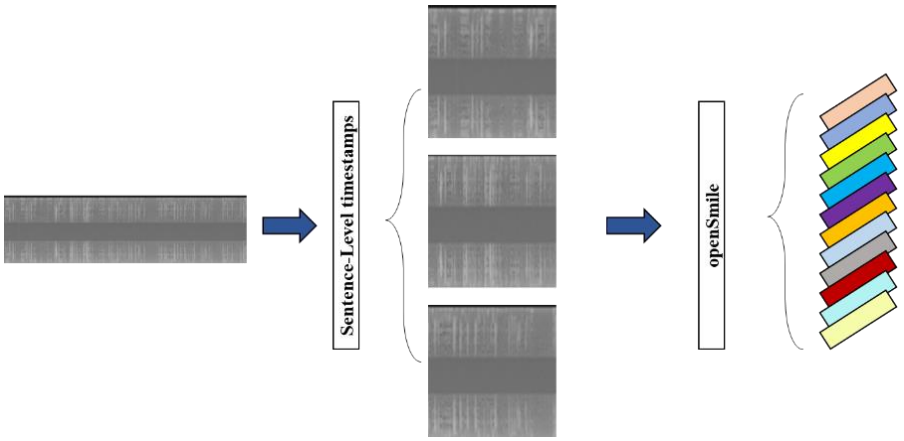

Figure 6. Audio features were extracted using the OpenSmile tool

3.3.3. Text Feature Extraction

The text data in this study is obtained through transcribing the interview audio of the patients. We include the speaker's name before each text utterance and segment the text based on periods. Furthermore, we make corrections to any transcribing errors to enhance the accuracy of feature extraction. In this experiment, we primarily employ EmoBertA [25] as the text feature extraction model. EmoBertA is trained based on RobertA, with additional training specifically focused on emotion recognition. By incorporating the speaker's name before each utterance and inserting separators between utterances in a conversation, EmoBertA can learn the states and contexts within and between speakers. Compared to RobertA, EmoBertA is trained on a larger emotion recognition dataset, resulting in superior performance in emotion recognition. Additionally, EmoBertA has demonstrated significant improvements over RobertA on various public datasets. Therefore, we have chosen this model for text feature extraction in this experiment.

<s> Docter: i was created to talk to people in a safe and secure environmenti
was created to talk to people in a safe and secure environment
Docter: how are you doing today
Patient: good
Docter: that's good. where are you from originally
Patient: atlanta Georgia
Docter: why'd you move to l_a
Patient: um my parents are from here um</s>

Figure 7. The BERT model is used to extract text features

4. EXPERIMENTS AND DISCUSSION

4.1. Experimental Environment and Parameter Settings

The experiments in this paper were conducted on a Linux operating system. The hardware setup for training primarily consisted of 8 GeForce RTX 3090 GPUs. The software environment utilized Python 3.10.18 and the PyTorch 1.10 framework. The experiment employed a Temporal Convolutional Network (TCN) as the underlying network architecture, which incorporated the gate mechanism of LSTM to capture long-term features. This was further optimized to obtain the Multi-modal Dual Text Contrastive Learning Model (MGTDCM). The main parameter settings for the model are as follows: an initial learning rate of 1e-3, Adam optimizer, 5 hidden layers and 12 hidden channels in the sentence-level feature extraction, a convolutional kernel size of 5, and a dropout rate of 0.3. Each experiment was conducted for a maximum of 100 epochs, with the learning rate decreasing by an order of magnitude every 5 epochs. If the loss on the test set did not decrease for 10 consecutive epochs, the training process was stopped. These settings were designed to effectively train and optimize the MGTDCM model, balancing model complexity and training efficiency.

4.2. Experimental Evaluation Metrics

In this study, we employed several metrics to evaluate the performance of different classification models, including Recall, Precision, F1-score, Specificity, and Accuracy.

Recall measures the proportion of actual positive samples that are correctly predicted as positive. A high recall indicates that the model is effective in identifying a greater number of depression patients, thereby reducing the rate of false negatives. Ideally, both Precision and Recall should be high for optimal model performance. However, there exists a trade-off between Precision and Recall, which can be expressed with the following formula for Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{20}$$

Precision measures the proportion of predicted positive samples that are actually positive. In the context of depression identification, a high precision indicates that the model can reduce false positives when diagnosing depression patients. This is crucial for avoiding unnecessary anxiety and subsequent medical interventions. The formula for precision is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{21}$$

The F1-score serves as the harmonic mean of Precision and Recall, taking into account both false positives and false negatives to provide a comprehensive evaluation of model performance. This metric is particularly useful when there is an uneven class distribution or when the costs of false positives and false negatives are different.

$$\text{F1} = \frac{2TP}{2TP + FP + FN} \tag{22}$$

Additionally, Specificity is an important metric that reflects the model's ability to identify healthy individuals, helping to ensure that those without depression are not incorrectly labeled as patients. This is crucial for minimizing the risk of misdiagnosis and unnecessary concern among individuals who are actually healthy. The formula for Specificity is as follows:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{23}$$

Lastly, Accuracy is a metric used to measure the proportion of correctly classified samples out of all samples in a classification model. It provides a general overview of the overall performance of the model, particularly when the class distribution is relatively balanced. The formula for Accuracy is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{24}$$

In the aforementioned formulas, TP refers to the number of true positive samples, which are correctly classified as positive. FP represents the number of false positive samples, which are incorrectly classified as positive. TN denotes the number of true negative samples, which are correctly classified as negative. FN represents the number of false negative samples, which are incorrectly classified as

negative. By considering these values, we can accurately assess the performance of a classification model.

4.3. Comparative Experiments On Different Dataset

In this section, we compared and evaluated different models on multimodal feature data. We primarily utilized the MDD2024 dataset and validated the models on the publicly available DAIC-WOZ dataset. To minimize any potential errors in the experiments, we selected the optimal values from three trials and took their average.

To assess the impact of different modalities on the models, we conducted two experiments on the MDD2024 dataset. This involved evaluating different unimodal features and combination methods. In the evaluation of different unimodal features, we employed separate unimodal models, treating video features, audio features, and text features as independent inputs. We compared the results against the reference TCN base model. When assessing various combinations of multimodal features, we employed a joint multimodal fusion attention mechanism to weight the features. Detailed experimental results can be found in Figure 8.

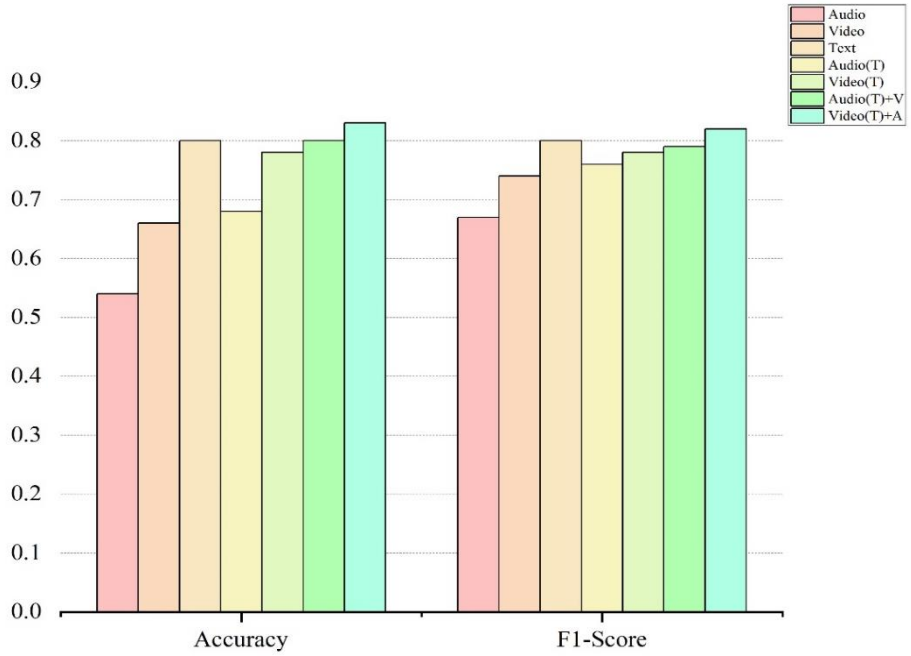

Figure 8. The results of depression recognition using different single-modal data and enhanced single-modal features with Joint Multi-modal Fusion Attention under the base model

In the figure, Audio(T) represents the audio features enhanced through transcribed text, while Video(T) represents the video features augmented by transcribed text. The experimental results indicate that both video and audio features are classified as weak features, whereas the transcribed text features are regarded as strong features. Subsequently, we utilized the transcribed text features as strong features and enhanced the two weak features individually before combining them. The results demonstrated that the combination of enhanced video features and audio features yielded the best performance. Therefore, in the subsequent experiments, we will consider video features as weak features and text features as strong features.

4.3.1. Result on self-collected clinical data

This section presents the experimental results obtained from the ongoing construction of the Chinese clinical dataset MDD2024. The results are summarized in Table 2. As evident from the table, a comparison between TCN(all) and TCN(sentence) reveals that using sentence-level data outperforms the direct use of complete paragraph data. Specifically, the introduction of sentence-level data led to

improvements in F1 score, specificity, and accuracy by 5%, 11%, and 7%, respectively, with the most significant increases observed in specificity and recall, which improved by 12% and 24%, respectively. Furthermore, a comparative experiment between TCN(all) and LSTM(all) shows that LSTM demonstrates superior performance in handling long-duration data, with a 3% increase in F1 score, an 8% increase in precision, and a 7% increase in accuracy. Finally, when comparing recent multimodal depression models with the proposed MGTDCM, it is evident that the MGTDCM model outperforms other models on the Chinese depression dataset MDD2024, achieving F1 score, specificity, precision, recall, and accuracy rates of 87%, 90%, 88%, 87%, and 87%, respectively. This indicates that the model exhibits excellent performance in the Chinese depression classification task, demonstrating commendable efficacy and generalization capabilities.

Table 2. The depression recognition results of our MGTDCM on MMD2024dataset

Model	F1-score	Specificity	Precision	Recall	Accuracy
TCN (all)	0.62	0.67	0.65	0.60	0.63
LSTM (all)	0.65	0.67	0.73	0.60	0.67
TCN (sentence)	0.67	0.78	0.77	0.84	0.70
BiAttention-GRU [11]	0.70	0.87	0.87	0.60	0.67
BiLSTM-GRU [26]	0.80	0.87	0.87	0.73	0.80
Zhuang et al. [27]	0.83	0.81	0.87	0.80	0.83
MGTDCM	0.87	0.90	0.88	0.87	0.87

In summary, the MGTDCM model demonstrates outstanding performance on real clinical data, effectively addressing practical issues and showcasing significant clinical applicability. It provides healthcare professionals with robust decision support and holds potential for application in the diagnosis and treatment of actual depression cases. By integrating multimodal data and leveraging advanced deep learning techniques, the MGTDCM model can offer more accurate and reliable depression classification results, significantly improving the diagnosis and treatment of depression patients. Therefore, the MGTDCM model holds immense potential for wide-ranging clinical applications.

4.3.2. Results on Public data

The experimental results presented in Table 3 are based on the publicly available DAIC-WOZ dataset. The Loss values are illustrated in Figure 9. The findings indicate that the proposed MGTDCM model outperforms in all evaluated metrics, achieving an F1 score, precision, and recall of 0.90. Additionally, the model attained peak values in precision, F1 score, and accuracy, underscoring its effectiveness and reliability in the task of depression classification.

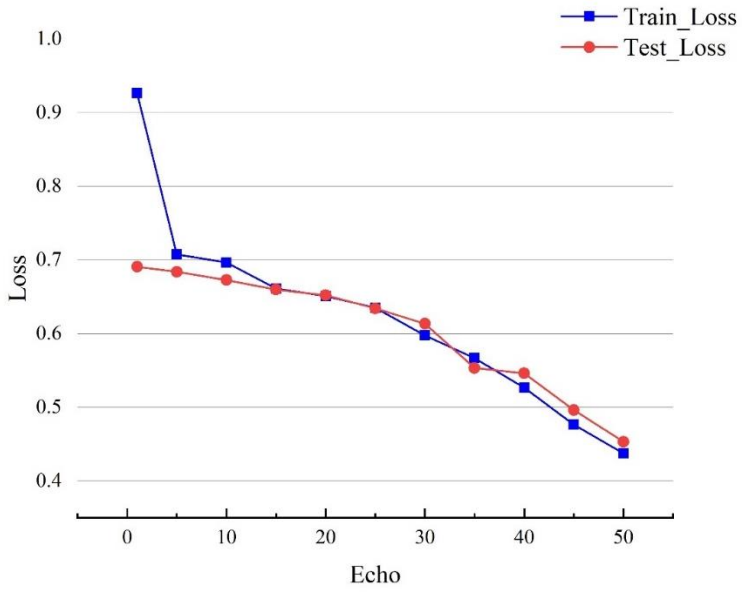

Figure 9. The TrainLoss and TestLoss values in this study

Table 3. The depression recognition results of our MGTDCM on DAIC-WOZ dataset

Model	F1-score	Specificity	Precision	Recall	Accuracy
Zhuang et al.	0.85	0.88	0.88	0.82	0.86
BiAttention-GRU	0.88	0.88	0.88	0.88	0.88
BiLSTM-GRU	0.85	0.79	0.79	0.92	0.86
Multi-modal LSTM	0.77	0.74	0.71	0.83	0.77
Zou et al. [28]	0.87	0.79	0.94	0.81	0.86
C-CNN	0.77	0.56	0.71	0.83	0.70
MGTDCM	0.90	0.88	0.90	0.90	0.89

In comparison, the Zhuang et al. model achieved an F1 score of 0.85, with moderate performance in terms of specificity and recall. The BiAttention-GRU model exhibited balanced performance across all metrics, with values of 0.88. The BiLSTM-GRU model demonstrated higher recall (0.92), but relatively lower specificity (0.79), which could potentially lead to an increase in false positive rates. Additionally, the overall performance of the Multi-modal LSTM and C-CNN models was weaker, with F1 scores of 0.77, indicating limitations in handling multimodal data. The Zou et al. model achieved an F1 score of 0.87, despite having high precision (0.94), the specificity remained relatively low (0.79). Overall, the MGTDCM model significantly improves the accuracy of depression detection by effectively integrating multimodal features, demonstrating its potential for application in this field.

4.4. Ablation

To further validate the superiority of the proposed model and the rationality of its additional components, we conducted ablation experiments on the publicly available DAIC-WOZ dataset. These experiments evaluated the performance of the model after adding different components while keeping other hyperparameters unchanged. The experimental results are shown in the table below:

Table 4. F1 score, specificity, precision, recall and accuracy metrics of the model after adding different components in the ablation experiment

Model	F1-score	Specificity	Precision	Recall	Accuracy
TCN (all)	0.77	0.75	0.74	0.81	0.78
TCN (sentence)	0.79	0.81	0.78	0.81	0.81
MDEFNet	0.80	0.82	0.80	0.82	0.82
MDEFNet+DTCL	0.83	0.84	0.83	0.83	0.84
MDEFNet+JMFATT	0.83	0.86	0.84	0.85	0.85
MGTDCM	0.90	0.88	0.90	0.90	0.89

To evaluate the impact of sentence-level structure on model performance, the TCN (all) model was initially employed as the baseline, which was subsequently transitioned to the TCN (sentence) model. The experimental results indicate that utilizing sentence-level data as input has a positive effect on model performance in the task of depression classification. In the following experiments, an LSTM model was integrated into the sentence-level TCN model to process global features. The results demonstrated a 1% increase in the F1 score, a 2% increase in precision, and a 1% increase in accuracy compared to the TCN (sentence) model. This outcome validates the significant performance enhancement of the proposed MDEFNet model on publicly available depression datasets, indicating that MDEFNet effectively strengthens various metrics in multimodal depression classification, including F1 score, precision, and accuracy. Furthermore, these findings further affirm the superiority and potential of the MDEFNet model in addressing tasks within public datasets. Subsequently, generated text was introduced as input within the MDEFNet model, incorporating the DTCL module to process dual-text features. Experimental results revealed improvements across all metrics based on the foundation of the MDEFNet model. Following this, the introduction of JMFATT aimed to amplify the expression of weaker modalities while capturing inter-modal similarities during the fusion process to enhance model accuracy. After integrating the attention mechanism, evaluation metrics improved by 3%, 4%, 4%, 3%, and 3% respectively, confirming the effectiveness of the joint multimodal fusion attention mechanism within the model. Finally, the combination of the dual-text contrastive learning module with the attention mechanism led to peak performance across all metrics. Compared to the baseline TCN model, the overall model performance saw significant enhancements, with metrics increasing by 13%, 13%, 16%, 9%, and 11%. Based on these experimental results, the MGTDCM model proposed in this study demonstrates exceptional performance in the task of depression classification, effectively integrating the advantages of its various components.

5. CONCLUSION

In this study, we propose the MGTDCM model, which combines the TCN model and LSTM in the field of depression classification. Addressing the challenges of limited depression datasets and model overfitting, the MGTDCM model incorporates multimodal text generation to identify depression in patients. The model generates text data using the ChatGLM model and combines it with other modal features to classify patients with depression. Experimental results demonstrate that the MGTDCM model improves performance across all evaluation metrics, both on the publicly available DAIC-WOZ dataset and the privately constructed MDD2024 dataset. The lack of original video files in publicly available datasets limits exploration in using multimodal data for depression identification, particularly in the visual modality. Additionally, since the existing datasets are mostly in English, differences in grammar and participants' living environments between English and Chinese could lead to variations in extracting text and audio features, potentially affecting the model's performance. Moreover, the limited number of participants and the unequal distribution of depression participants in the dataset pose challenges and influence the effectiveness of the model, requiring additional modeling efforts. Therefore, we are currently constructing the MDD2024 Chinese depression

database, which will have a larger number of participants, include more visual feature files, and incorporate multiple scale data evaluated by professional psychiatrists. This will further facilitate research in the field of Chinese depression identification. To protect the privacy of participants' personal health information, the MDD2024 dataset will not include any information that compromises participants' privacy, including original video files. Finally, although the MGTDCM model is currently a binary classification model and cannot identify different levels of depression, future work will involve incorporating semantic emotions, continuously optimizing feature information in the MDD2024 dataset, and exploring more effective multi-classification depression classification models to enhance the accuracy and practicality of depression classification.

ACKNOWLEDGEMENTS

Shanxi Province Natural Science Foundation: Research on Core Scalable and Assurance Intelligent Service Model for Cross-Network Virtual Reality (201801D121147)

REFERENCES

[1] WOODY C, FERRARI A, SISKIND D, et al. A systematic review and meta-regression of the prevalence and incidence of perinatal depression [J]. 2017, 219: 86-92.

[2] SANTOMAURO D F, HERRERA A M M, SHADID J, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic [J]. 2021, 398(10312): 1700-12.

[3] CD M J P M. Projections of global mortality and burden of disease from 2002 to 2030 [J]. 2006, 3: 2011-30.

[4] EVANS-LACKO S, AGUILAR-GAXIOLA S, AL-HAMZAWI A, et al. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO World Mental Health (WMH) surveys [J]. 2018, 48(9): 1560-71.

[5] WANG Q, YANG H, YU Y J J O V C, et al. Facial expression video analysis for depression detection in Chinese patients [J]. 2018, 57: 228-33.

[6] SENN S, TLACHAC M, FLORES R, et al. Ensembles of bert for depression classification; proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), F, 2022 [C]. IEEE.

[7] FLORES R, TLACHAC M, TOTO E, et al. Transfer learning for depression screening from follow-up clinical interview questions [M]. Deep Learning Applications, Volume 4. Springer. 2022: 53-78.

[8] WANG J, RAVI V, ALWAN A. Non-uniform speaker disentanglement for depression detection from raw speech signals; proceedings of the Interspeech, F, 2023 [C]. NIH Public Access.

[9] RAY A, KUMAR S, REDDY R, et al. Multi-level attenstion network using text, audio and video for depression prediction; proceedings of the Proceedings of the 9th international on audio/visual emotion challenge and workshop, F, 2019 [C].

[10] RODRIGUES MAKIUCHI M, WARNITA T, UTO K, et al. Multimodal fusion of bert-cnn and gated cnn representations for depression detection; proceedings of the Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, F, 2019 [C].

[11] CAO Y, HAO Y, LI B, et al. Depression prediction based on BiAttention-GRU [J]. 2022, 13(11): 5269-77.

[12] BUCUR A-M, COSMA A, ROSSO P, et al. It's just a matter of time: Detecting depression with time-enriched multimodal transformers; proceedings of the European Conference on Information Retrieval, F, 2023 [C]. Springer.

[13] RAHMAN A B S, TA H-T, NAJJAR L, et al. DepressionEmo: A novel dataset for multilabel classification of depression emotions [J]. 2024, 366: 445-58.

[14] GIMENO-GóMEZ D, BUCUR A-M, COSMA A, et al. Reading Between the Frames: Multi-modal Depression Detection in Videos from Non-verbal Cues; proceedings of the European Conference on Information Retrieval, F, 2024 [C]. Springer.

[15] MCGINNIS E W, ANDERAU S P, HRUSCHAK J, et al. Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood [J]. 2019, 23(6): 2294-301.

[16] XU X, WANG Y, WEI X, et al. Attention-based acoustic feature fusion network for depression detection [J]. 2024, 601: 128209.

[17] HAQUE A, GUO M, MINER A S, et al. Measuring depression symptom severity from spoken language and 3D facial expressions [J]. 2018.

[18] LIU Y, OTT M, GOYAL N J A P A. Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach [J]. 2019, 1(3.1): 3.

[19] SERMANET P, LYNCH C, CHEBOTAR Y, et al. Time-contrastive networks: Self-supervised learning from video; proceedings of the 2018 IEEE international conference on robotics and automation (ICRA), F, 2018 [C]. IEEE.

[20] YU Y, SI X, HU C, et al. A review of recurrent neural networks: LSTM cells and network architectures [J]. 2019, 31(7): 1235-70.

[21] GLM T, ZENG A, XU B, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools [J]. 2024.

[22] KROENKE K, STRINE T W, SPITZER R L, et al. The PHQ-8 as a measure of current depression in the general population [J]. 2009, 114(1-3): 163-73.

[23] AMOS B, LUDWICZUK B, SATYANARAYANAN M J C S O C S. Openface: A general-purpose face recognition library with mobile applications [J]. 2016, 6(2): 20.

[24] EYBEN F, WöLLMER M, SCHULLER B. Opensmile: the munich versatile and fast open-source audio feature extractor; proceedings of the Proceedings of the 18th ACM international conference on Multimedia, F, 2010 [C].

[25] KIM T, VOSSEN P J A P A. Emoberta: Speaker-aware emotion recognition in conversation with roberta [J]. 2021.

[26] SHEN Y, YANG H, LIN L. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model; proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), F, 2022 [C]. IEEE.

[27] CHEN Z, DENG J, ZHOU J, et al. Depression detection in clinical interviews with LLM-empowered structural element graph; proceedings of the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), F, 2024 [C].

[28] ZOU B, HAN J, WANG Y, et al. Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders [J]. 2022, 14(4): 2823-38.