Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 14, No. 1, 2025

Multimodal Depression Recognition Based on Sentence-level Dynamic Multimodal Split Attention Fusion

Mingyang Sun [1], Shukui Ma [2], Guangping Zhuo [1], Fei Ma [3] and Guanghua Zhang [3, *]

[1] School of Computer Science and Technology, Taiyuan Normal University, Jinzhong Shanxi, 030619, China
[2] Department of information, the 985 Hospital of Logistics Support Forces of Chinese PLA, Taiyuan Shanxi, 030001, China
[3] Department of Computer Science and Technology, Taiyuan University, Taiyuan Shanxi, 030032, China

*** Corresponding author:** Guanghua Zhang

Abstract: Depression is a common yet highly covert mental disorder, making the development of efficient intelligent recognition methods crucial for early screening and clinical diagnostic support. Existing multimodal depression recognition approaches still face limitations in modal interaction and long-sequence semantic modeling, struggling to fully capture local dynamics and cross-modal dependencies. To address this, this study proposes a multimodal temporal fusion network. This approach first divides long medical interview sequences into sentence-level units based on timestamps to mitigate information dilution in lengthy sequences. Subsequently, it designs a sentence-level dynamic multimodal attention fusion module. This module further segments sentence sequences into contiguous segments and adaptively emphasizes key modal features while suppressing redundant and noisy information through dynamic weight allocation. On the public dataset DAIC-WOZ and the self-built Chinese dataset MDD2025, MTFNet achieves accuracy rates of 86% and 84%, respectively.

Keywords: Depression; Multimodal; Sentence-level Dynamic Multimodal Split Attention Fusion.

1. Introduction

Depression is a highly prevalent mental disorder affecting approximately 350 million people worldwide, causing severe impairment in personal functioning and significant societal burden, particularly in low- and middle-income countries[1]. As one of the primary contributors to disability-adjusted life years (DALYs) lost[2],[3], the World Health Organization notes that depression's disease burden ranking is rapidly rising, projected to become the world's second-leading health issue by 2030[4],[5]. Depression exhibits high prevalence across all age groups, affecting approximately 4.5% of adolescents[6] , with postpartum depression occurring in about 20% of mothers within three months of childbirth[7],[8]. Among the elderly population, prevalence ranges from 10% to 20%[9]. Challenges persist in early identification and intervention due to factors such as uneven distribution of medical resources, prolonged diagnosis cycles, and limitations of subjective assessments.

Early studies predominantly employed unimodal modeling. For speech, Kim et al.[10] utilized CNNs to model log-Mel spectrograms of Korean speech, validating mobile depression screening feasibility. For text, Amanat et al.[11] employed a two-layer LSTM/RNN to identify depressive expressions on social platforms. For vision, Zhang et al.[12] constructed an interpretable model from facial action units. While unimodal approaches achieved some success, their performance remains limited due to noise, information gaps, and environmental factors. Consequently, multimodal fusion has emerged as the mainstream approach. Zhang et al.[13] proposed SMFL to enhance weak modal features through cross-modal attention; Wang et al.[14] designed an audio-text fusion model based on SESE, achieving promising results.

Despite advances in multimodal depression detection, challenges persist, including insufficient intermodal interaction and difficulties in representing long-sequence semantics. Differences in temporal dynamics and expressive forms across modalities hinder models from capturing cross-modal dependencies, while limited high-quality data often leads to overfitting. To address these challenges, this study proposes a multimodal temporal fusion network. It segments sequences into sentence units and employs hierarchical modeling with LSTM and BiLSTM to enhance long-range semantic representation. Additionally, it designs a sentence-level Dynamic Multimodal Split Attention Fusion Module that utilizes segmented attention and learnable weights to amplify key modalities and suppress noise.

2. Model Construction

To address insufficient modality interaction and challenges in modeling long-sequence semantics in multimodal depression recognition, this study proposes the Multimodal Temporal Fusion Network (MTFNet), as shown in Fig. 1. This model takes consultation data from three modalities—text, audio, and facial expressions—as input. It sequentially undergoes sentence-level feature modeling, dynamic attention fusion, and global temporal modeling to achieve multi-level feature representation and fusion.

At the sentence-level modeling stage, multimodal data is aligned to sentence-level units based on timestamps and fed into LSTMs for feature extraction. Subsequently, a sentence-level dynamic multimodal segmentation attention fusion module segments features into blocks. Through dual weighting by block-level attention and modality weights, key features are amplified while noise is suppressed, addressing the limitation of traditional concatenation in capturing fine-grained interactions. During the global temporal modeling phase, audio and facial features are fed into an LSTM, while textual features enter a BiLSTM to capture cross-sentence dependencies, forming a global discourse representation.

Finally, the fused features undergo a fully connected layer to perform depression state classification.

2.1. Foundational Model Overview

Multimodal data exhibits significant temporal dependencies, making sequential modeling crucial for depression detection. LSTM[15] mitigates gradient vanishing through gating mechanisms, preserving key information over extended time spans and enabling dynamic modeling of speech and facial modalities. BiLSTM[16] integrates forward and backward information to model context more comprehensively, making it particularly suitable for context-sensitive text modalities. Based on this, MTFNet employs LSTM for speech and facial modalities while utilizing BiLSTM for global temporal modeling in text modalities, thereby balancing local dynamics and global dependencies.

2.2. Sentence-Level Dynamic Multimodal Split Attention Fusion Module

Existing methods often fuse sentence sequences as a whole, struggling to capture fragment dynamics and modality differences. This study improves upon MSAF[17] by proposing the Sentence-level Dynamic Multimodal Split Attention Fusion (SDMSAF). This module incorporates sentence sequence segmentation processing and dynamic multimodal attention fusion, enabling local fragment modeling and adaptive modal weight allocation to enhance expressive capabilities in interactions.

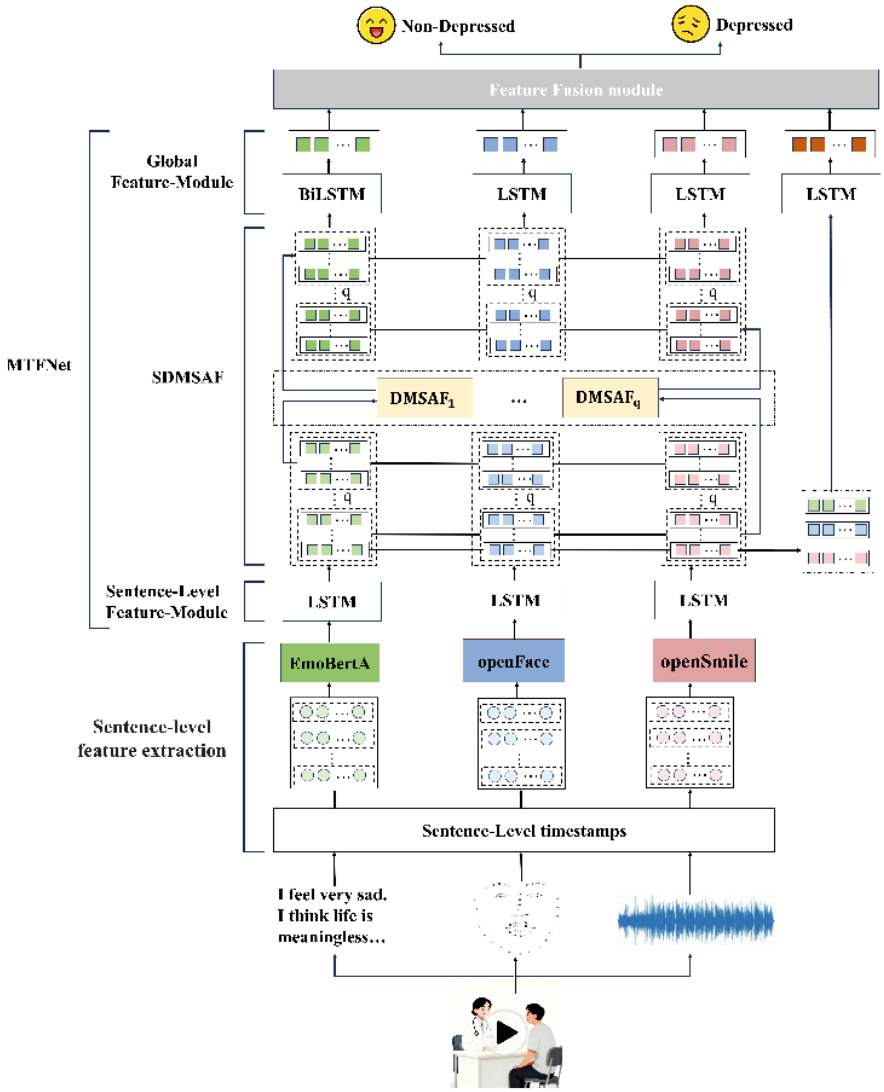

Fig 1. Multimodal Temporal Fusion Network

2.2.1. Sentence Sequence Segmentation Mechanism

The MSAF module was originally designed for convolutional neural networks, with its core idea being to segment features into blocks and assign differentiated attention. However, when directly applied to sequence modeling, it lacks flexible adjustment along the temporal dimension. Inputting the entire sequence into DMSAF reduces computation but struggles to capture temporal dynamics; inputting sentence-by-sentence enhances local modeling but is prone to overfitting. This study proposes the Sentence-Sequence Segmentation (SSS) mechanism, illustrated in Fig. 2. By partitioning sentence sequences into contiguous segments, this mechanism enables DMSAF to capture dynamic information within local scopes while balancing model complexity and representational power through controlled segment counts.

After initial sentence-level feature extraction, the resulting sentence sequence representations across audio, facial, and text modalities are expressed as in Equations (1–3):

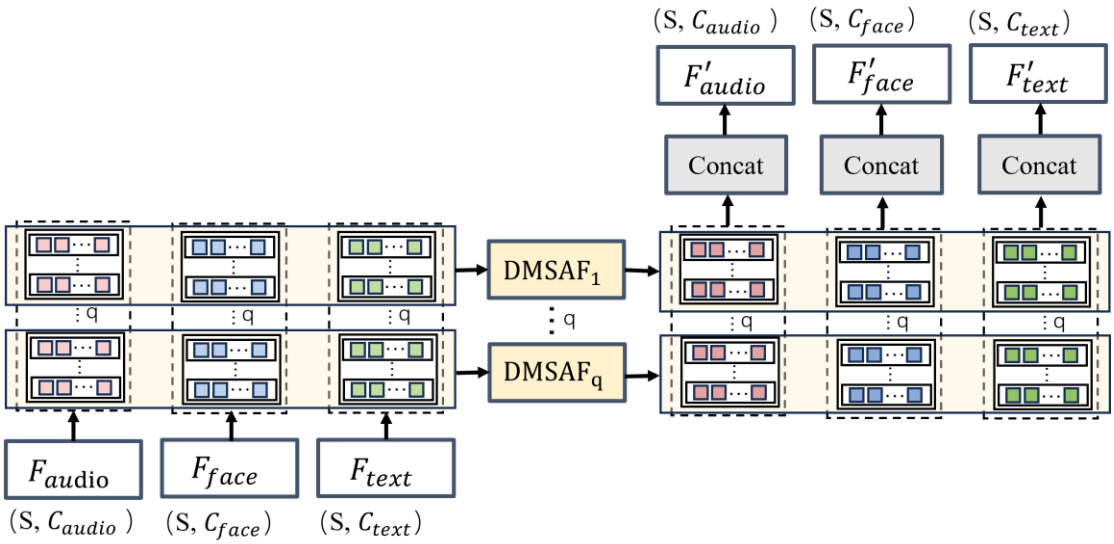

Fig 2. Sentence-Sequence Segmentation Mechanism

$$F^{(text)} = \left\{f_1^{(text)}, f_2^{(text)}, \dots, f_S^{(text)}\right\} \tag{1}$$

$$F^{(audio)} = \left\{f_1^{(audio)}, f_2^{(audio)}, \dots, f_S^{(audio)}\right\} \tag{2}$$

$$F^{(face)} = \left\{f_1^{(face)}, f_2^{(face)}, \dots, f_S^{(face)}\right\} \tag{3}$$

Where S denotes the total number of sentences. Subsequently, each modality sequence is segmented into q fragments along the sentence sequence dimension, as shown in Equations (4–6):

$$F^{(text)} = \left[F_1^{(text)}, F_2^{(text)}, \dots, F_q^{(text)}\right] \tag{4}$$

$$F^{(audio)} = \left[F_1^{(audio)}, F_2^{(audio)}, \dots, F_q^{(audio)}\right] \tag{5}$$

$$F^{(face)} = \left[F_1^{(face)}, F_2^{(face)}, \dots, F_q^{(face)}\right] \tag{6}$$

Among these, $F_i^{(\text{text})}$, $F_i^{(\text{audio})}$, $F_i^{(\text{face})}$ represent the i th sentence fragment sub-sequence for text, audio, and face modalities respectively. Each modality is divided into q sub-sequence fragments along the sentence sequence dimension, with each fragment having a length of $s = \left\lfloor \frac{S}{q} \right\rfloor$, C_{text}、C_{audio}、C_{face} denote the sentence feature dimensions for each modality. Finally, the multimodal data aligned under the same fragment index i are jointly input into a single $DMSAF_i$, as shown in Equation (7):

$$H_i = DMSAF_i\left(F_i^{(\text{text})}, F_i^{(\text{audio})}, F_i^{(\text{face})}\right) \tag{7}$$

2.2.2. Dynamic Multimodal Split Attention Fusion Module

Existing research indicates that text, audio, and visual modalities contribute differently to depression recognition, yet traditional fusion methods often treat them equally, overlooking these distinctions. To address this, this study proposes the Dynamic Multimodal Split Attention Fusion (DMSAF) module, as illustrated in Fig. 3. The DMSAF module adaptively evaluates the importance of each modality through learnable modality weights, enhancing key modalities and suppressing redundant noise during fusion to improve discriminative capability.

For the input three-modal feature maps $F_i^{(\text{text})}, F_i^{(\text{audio})}, F_i^{(\text{face})}$, uniformly denoted as F_m in this section, where $m \in \{\text{text,audio,face}\}$ represents the modal type. First, the features are partitioned along the channel dimension, dividing each modal feature map into several sub-blocks. If the number of channels cannot be evenly divided, zero padding is applied at the end. As shown in Equation (8):

Here, B_m^j denotes the j th channel block of modality m, and $|B_m|$ represents the number of sub-blocks. To aggregate cross-modality contextual information, we first perform element-wise summation across all channel blocks within each modality, yielding the modality-level summary matrix S_m as shown in Equation (9). Subsequently, global average pooling is applied at the sentence dimension to obtain the modality descriptor D_m as shown in Equation (10).

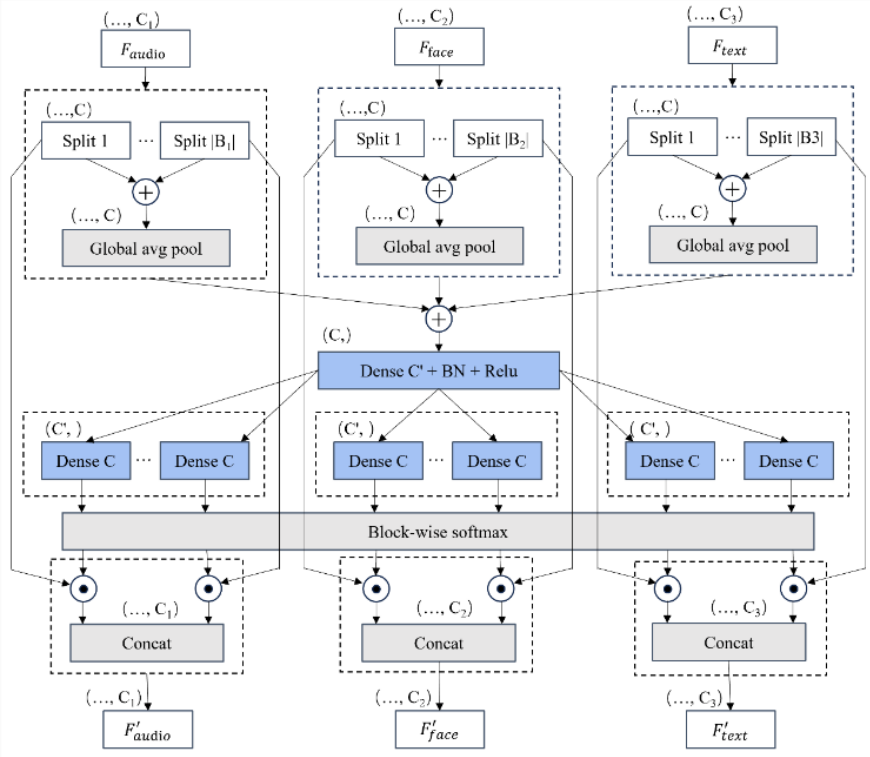

Fig 3. Dynamic Multimodal Split Attention Fusion Module

$$F_m = \left[B_m^1, B_m^2, \ldots, B_m^{|B_m|}\right] \tag{8}$$

$$S_m = \sum_{j=1}^{|B_m|} B_m^j \tag{9}$$

$$D_m(c) = \frac{1}{S}\sum_{n=1}^{S} S_m(n,c) \tag{10}$$

Here, c denotes the channel index, and n denotes the sentence index. Subsequently, DMSAF assigns learnable parameters β_m to each modality and normalizes them via softmax to obtain modality importance coefficients α_m, as shown in Equation (11). Finally, the cross-modal joint description representation G is obtained, as shown in Equation (12):

$$\alpha_m = \frac{\exp(\beta_m)}{\sum_{M\in\text{text,audio,face}} \exp(\beta_M)} \tag{11}$$

$$G = \alpha_{\text{text}} D_{\text{text}} + \alpha_{\text{audio}} D_{\text{audio}} + \alpha_{\text{face}} D_{\text{face}} \tag{12}$$

The weighted fusion of global representations G is subsequently fed into a fully connected layer for nonlinear transformation, yielding the reduced-dimensional joint representation Z as shown in Equation (13):

$$Z = \text{ReLU}\big(\text{BN}(W_Z G + b_Z)\big) \tag{13}$$

Where $W_Z \in \mathbb{R}^{C'\times C}, C' = \left\lfloor \frac{C}{r} \right\rfloor$ and r represent the dimensionality reduction factor. This operation reduces redundant information and parameter size by compressing feature dimensions, thereby mitigating overfitting and improving computational efficiency.

Based on the compressed representation Z, for each channel block B_m^j, compute its original score vector $U_m^j \in R^C$ as shown in Equation (14), and then use softmax to obtain the block-level attention weights A_m^j as shown in Equation (15):

$$U_m^j = W_m^j Z + b_m^j \tag{14}$$

$$A_m^j = \frac{\exp(U_m^j)}{\sum_k \sum_l \exp(U_k^l)} \tag{15}$$

Where W_m^j, b_m^j represents learnable parameters. To prevent excessive suppression of low-weight blocks, a regularization factor $\lambda \in [0,1]$ is introduced. The optimized channel block representation is shown in Equation (16).

Finally, all optimized channel blocks are concatenated along the channel dimension to obtain the optimized modal feature $\hat{F}_{text}$、$\hat{F}_{audio}$、$\hat{F}_{face}$, as shown in Equation (17):

$$\hat{B}_m^j = [\lambda + (1-\lambda) \cdot A_m^j] \odot B_m^j \tag{16}$$

$$\hat{F}_m = \left[\hat{B}_m^1, \hat{B}_m^2, \cdots, \hat{B}_m^{|B_m|}\right] \tag{17}$$

3. Experimental Data

3.1. Datasets

This study utilizes the Chinese depression dataset MDD2025, constructed in collaboration with local hospitals, and incorporates the public dataset DAIC-WOZ for comparative validation. MDD2025 was collected through clinical interviews by professional psychiatrists, with patients completing consultations in private rooms after signing informed consent forms. After excluding invalid data—including videos shorter than three minutes, missing scale scores, or cases without detectable facial information—the dataset comprised 522 valid samples: 212 males and 310 females. DAIC-WOZ is an international public dataset containing audio, video, transcribed text, and scale scores. This study classified participants with PHQ-8 scores ≥9 as the depression group and the remainder as the non-depression group, based on the PHQ-8 scale[18] . The depression group comprised 334 individuals, and the non-depression group comprised 188 individuals.

3.2. Data Preprocessing

The MDD2025 and DAIC-WOZ datasets comprise 522 and 192 samples respectively, with 334 and 74 depression-positive cases. Both datasets were split into training and testing sets at an 8:2 ratio to maintain class balance. To mitigate class imbalance, class weights were introduced using the formula in (18):

$$\text{weight} = \left[\frac{S_t}{N_0}, \frac{S_t}{N_1}\right] \tag{18}$$

where S_t denotes the total number of samples in the training set; N_0 and N_1 represent the number of negative and positive samples in the training set, respectively.

3.3. Multimodal Feature Extraction

3.3.1. Temporal Facial Feature Extraction

To ensure temporal alignment between facial features and speech/text modalities, the raw video is segmented by sentence-level timestamps and standardized, while frames lacking speech activity or containing occlusions are discarded. OpenFace is employed for frame-level analysis of video segments, extracting multidimensional temporal features including head pose, eye gaze direction, facial action units (AUs), and coordinates of 68 facial landmarks. A time step of 0.033 seconds is adopted to enhance feature extraction accuracy and efficiency.

3.3.2. Audio Feature Extraction

Audio data is segmented into speech clips based on sentence-level timestamps. openSMILE is employed to extract emotion-related acoustic features, utilizing the eGeMAPS feature set. This includes fundamental frequency (F0), spectral energy, MFCC, pitch variation, and loudness dynamics. These features reflect abnormal patterns in depressed patients, such as slowed speech rate, flat intonation, and reduced energy, providing valuable auxiliary information for the model to discern depressive states.

3.3.3. Text Feature Extraction

Text data is derived from manually transcribed audio recordings of medical consultations. During preprocessing, noise elements such as filler words, non-linguistic markers, and garbled characters are removed to ensure logical, structural, and semantic accuracy of the text, providing standardized and reliable input for the model. This study employs EmoBertA for text feature extraction. By prefixing speaker names to utterances and inserting delimiters between dialogue segments, the model captures both speaker-internal states and inter-speaker contextual relationships.

4. Experiments and Discussion

4.1. Experimental Environment and Parameter Settings

This study's experiments were conducted on the Linux operating system. The training hardware included three NVIDIA GeForce RTX 3080 GPUs, with the software environment utilizing Python 3.7.12 and the PyTorch 1.8.0 framework. Key model parameters are configured as follows: The LSTM hidden layer dimension for sentence-level feature extraction is set to 16. The number of hidden units for inter-sentence temporal modeling LSTMs is 8 (audio and face modalities) and 4 (text modality, bidirectional structure). Dropout is set to 0.3. The Adam optimizer is employed with an initial learning rate of 1e-3.

4.2. Experimental Evaluation Metrics

To comprehensively evaluate the performance of various classification models in depression recognition tasks, this study employs five commonly used metrics: Recall, Precision, F1-score, Specificity, and Accuracy, as shown in Equations (19–22).

Recall measures a model's ability to identify genuinely

depressed patients; higher values indicate lower false negative rates. Precision measures the proportion of samples predicted as depressed that are actually depressed, with higher precision reducing false positive rates; F1-score is the harmonic mean of precision and recall, reflecting the combined performance of false positive and false negative control, suitable for imbalanced data; Accuracy indicates the overall classification correctness rate across all samples, but may mask minority class bias in imbalanced scenarios, thus requiring comprehensive analysis with other metrics.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{19}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{20}$$

$$\text{F1score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

Here, TP denotes the number of correctly classified positive samples; FP denotes the number of negative samples misclassified as positive; TN denotes the number of correctly classified negative samples; FN denotes the number of positive samples misclassified as negative.

4.3. Experimental Results and Analysis

4.3.1. Experimental Results on the MDD2025 Dataset

On the MDD2025 dataset, this study compares the performance of the proposed MTFNet model against multiple models, with results shown in Table 1. Traditional full-sequence modeling approaches like TCN and LSTM lack the ability to capture fine-grained local features, resulting in limited performance. Sentence-level TCN and LSTM models, which partition sequences into sentence-level units for modeling, capture local dynamics more effectively, leading to significant performance improvements. BiLSTM-GRU further outperforms sentence-level LSTM, while the proposed MTFNet achieves the best overall performance, surpassing all comparison models across metrics. This fully validates its effectiveness in multimodal depression recognition.

4.3.2. Experimental Results on the DAIC-WOZ Dataset

Comparative results on the public DAIC-WOZ dataset are shown in Table 2. The MTFNet model achieves the best performance on the DAIC-WOZ dataset, with an F1-score of 0.85, accuracy of 0.86, specificity and precision of 0.87 and 0.86 respectively, and recall of 0.84.

4.4. Module Ablation Experiments

To further validate the effectiveness of the sentence-level segmentation and multimodal dynamic fusion mechanism in the proposed MTFNet model, ablation experiments were conducted on the DAIC-WOZ dataset. The results are shown in Table 3. After introducing the Dynamic Multimodal Split Attention Fusion Module, using only this module without sequence segmentation (q=1) improved the F1-score and accuracy to 0.81 and 0.82, respectively, indicating that cross-modal dynamic fusion significantly enhances the model's representational capability. Experiments were conducted with varying segmentation levels (q) to investigate the impact of sentence segmentation on model performance. Results show that both q=3 and q=5 outperform q=1, indicating that moderate segmentation aids in capturing intra-sequence dynamics. Optimal performance is achieved at q=4, balancing local information capture with global context. Performance slightly declines at q=5 due to overfitting in limited samples caused by excessive segmentation increasing parameters. In summary, sentence-level segmentation is essential for multimodal sequence modeling, and segment count selection is critical. MTFNet achieves optimal performance at q=4, validating the effectiveness and superiority of segmentation processing combined with dynamic multimodal attention fusion.

Table 1. Depression classification results of the MTFNet model on the MMD2025 dataset

Model	F1score	Precision	Recall	Accuracy
TCN (all)	0.61	0.62	0.60	0.63
LSTM (all)	0.64	0.66	0.62	0.65
TCN (sentence)	0.67	0.71	0.63	0.65
LSTM (sentence)	0.69	0.75	0.65	0.71
BiLSTM-GRU[19]	0.79	0.81	0.78	0.80
MTFNet	0.83	0.84	0.83	0.84

Table 2. Depression classification results of the MTFNet model on the DAIC-WOZ dataset

Model	F1score	Precision	Recall	Accuracy
TCN (all)	0.68	0.71	0.66	0.70
LSTM (all)	0.71	0.78	0.66	0.72
TCN (sentence)	0.72	0.75	0.68	0.74
LSTM (sentence)	0.74	0.78	0.70	0.76
BiLSTM-GRU[19]	0.83	0.84	0.82	0.84
MTFNet	0.85	0.86	0.84	0.86

Table 3. Ablation Experiment Results

Model	F1score	Precision	Recall	Accuracy
TCN (all)	0.74	0.78	0.70	0.76
LSTM (all)	0.81	0.82	0.80	0.82
TCN (sentence)	0.83	0.84	0.82	0.84
LSTM (sentence)	0.82	0.83	0.81	0.83
MTFNet	0.85	0.86	0.84	0.86

5. Conclusion

This paper addresses the challenges of insufficient semantic representation in long sequences and limited modal interaction in multimodal depression recognition by proposing the Multimodal Temporal Fusion Network (MTFNet). The model takes text, audio, and facial landmarks as inputs. It first segments long interview sequences into sentence-level units and performs preliminary alignment by extracting sentence-level features from each modality via LSTMs. Subsequently, a sentence-level Dynamic Multimodal Split Attention Fusion Module is introduced. This module employs segmented attention and learnable modality weights across consecutive segments to achieve cross-modal fine-grained interaction and dynamic importance allocation. Finally, it employs BiLSTM (text) and LSTM (audio, facial) for global temporal modeling and classification. Experimental results demonstrate that MTFNet outperforms TCN, LSTM, and existing comparison models on both the MDD2025 and DAIC-WOZ datasets. Ablation studies validate the effectiveness of segmented processing and dynamic attention, providing an efficient multimodal fusion solution for automatic depression detection.

Acknowledgments

This research was supported by the Postgraduate Practical Innovation Project of Taiyuan Normal University (Grant No. SYYJSYC-2580).

References

[1] Depression WHO. Other common mental disorders: global health estimates[J]. Geneva: World Health Organization, 2017, 24(1).

[2] Wittchen H U, Jacobi F, Rehm J, et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010[J]. European neuropsychopharmacology, 2011, 21(9): 655-679.

[3] Lim S S, Vos T, Flaxman A D, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010[J]. The lancet, 2012, 380(9859): 2224-2260.

[4] Mathers C D, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030[J]. PLoS medicine, 2006, 3(11): e442.

[5] World Health Organization. The global burden of disease: 2004 update [M]. Geneva: World Health Organization, 2008.

[6] Li J Y, Li J, Liang J H, et al. Depressive symptoms among children and adolescents in China: a systematic review and meta-analysis[J]. Medical Science Monitor, 2019, 25: 7459–7470. DOI: 10.12659/MSM.916774.

[7] Marcus S M, Flynn H A, Blow F C, et al. Depressive symptoms among pregnant women screened in obstetrics settings[J]. Journal of Women's Health, 2003, 12(4): 373–380. DOI: 10.1089/154099003765448886.

[8] Grace S L, Evindar A, Stewart D E. The effect of postpartum depression on child cognitive development and behavior: A review and critical analysis of the literature[J]. Archives of Women's Mental Health, 2003, 6(4): 263–274. DOI: 10.1007/s00737-003-0024-6.

[9] Zenebe Y, Akele B, W/Selassie M, Necho M. Prevalence and determinants of depression among old age: a systematic review and meta-analysis[J]. Annals of General Psychiatry, 2021, 20: 55. DOI: 10.1186/s12991-021-00375-x.

[10] Kim A Y, Jang E H, Lee S H, et al. Automatic depression detection using smartphone-based text-dependent speech signals: deep convolutional neural network approach[J]. Journal of medical Internet research, 2023, 25: e34474.

[11] Amanat A, Rizwan M, Javed A R, et al. Deep learning for depression detection from textual data[J]. Electronics, 2022, 11(5): 676.

[12] Mahayossanunt Y, Nupairoj N, Hemrungrojn S, et al. Explainable depression detection based on facial expression using LSTM on attentional intermediate feature fusion with label Smoothing[J]. Sensors, 2023, 23(23): 9402.

[13] Zhang G, Zhuo G, Yang Y, et al. Sentence-level multi-modal feature learning for depression recognition[J]. Frontiers in Psychiatry, 2025, 16: 1439577.

[14] Ye J, Yu Y, Wang Q, et al. Multi-modal depression detection based on emotional audio and evaluation text[J]. Journal of Affective Disorders, 2021, 295: 904-913.

[15] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[16] Liu G, Guo J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification[J]. Neurocomputing, 2019, 337: 325-338.

[17] Lang S, Chuqing H, Guofa L, et al. MSAF: multimodal split attention fusion[J]. CoRR, 2020.

[18] A K K , B T W S , C R L S ,et al. The PHQ-8 as a measure of current depression in the general population[J].Journal of Affective Disorders, 2009, 114(1–3):163-173.DOI: 10. 1016/ j.jad.2008.06.026.

[19] 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.