Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 6, No. 2, 2026

A Lightweight Multimodal Feature Alignment Framework for Depression Detection

Jiangfeng Liu

College of Electronics and Information, Southwest Minzu University, Chengdu, China

Abstract: With the rapid advancement of artificial intelligence, computer vision, speech recognition, and natural language processing, automatic depression detection based on multimodal data has attracted increasing research attention. Compared with unimodal approaches, multimodal fusion leverages complementary information from speech, text, facial expressions, and other behavioral cues, thereby improving the accuracy and robustness of depression assessment. However, existing multimodal models are often computationally intensive and parameter-heavy, which limits their deployment on resource-constrained devices. In addition, semantic and distributional discrepancies across modalities pose significant challenges for effective feature alignment, adversely affecting fusion performance. To address these issues, this paper proposes a lightweight multimodal feature alignment framework for depression severity estimation. The proposed method constructs lightweight feature extraction networks and introduces a cross-modal feature alignment mechanism to enable effective mapping and fusion across heterogeneous feature spaces. While significantly reducing model size and computational complexity, the framework maintains competitive predictive performance. Experimental results on multiple public depression datasets demonstrate that the proposed approach achieves a mean absolute error (MAE) of 4.44 and a root mean square error (RMSE) of 5.77 in PHQ-8 score estimation, indicating strong generalization capability and practical deployment potential.

Keywords: Multimodal learning, Feature alignment, Lightweight model, Depression severity estimation, PHQ-8.

1. Introduction

Depression is a common and severe mental disorder characterized by persistent low mood, loss of interest, cognitive impairment, and sleep disturbances. According to the World Health Organization (WHO), depression has become one of the leading causes of disability worldwide, imposing significant burdens on individuals and society. Traditional diagnosis of depression primarily relies on clinical interviews and self-report scales, which are subjective, time-consuming, and unsuitable for large-scale screening. Automatic depression assessment using behavioral signals such as speech has therefore attracted increasing research interest [1].

In recent years, with the rapid advancement of deep learning technologies, automatic depression recognition based on multimodal data has attracted increasing attention. Multimodal data typically include speech signals, facial expression videos, textual content, and physiological signals, which reflect emotional and psychological states from different perspectives. For example, individuals with depression may exhibit monotonic intonation and slower speech rate in audio, reduced facial expressiveness in video, and negative sentiment tendencies in textual expression. Recent studies have explored multimodal modeling of audio and textual signals for automatic depression detection and severity estimation [2].

Compared with unimodal approaches, multimodal methods can integrate complementary information from multiple sources, thereby improving robustness and accuracy. However, multimodal learning faces two major challenges: (1) high model complexity, which limits real-world deployment, and (2) distributional discrepancies across modalities, leading to difficulties in cross-modal feature alignment and suboptimal fusion performance. As discussed in prior studies on multimodal machine learning, heterogeneity and representation gaps across modalities remain fundamental challenges for effective fusion [3]. Therefore, developing a lightweight and effective multimodal feature alignment method is of significant theoretical and practical importance.

This study investigates cross-modal semantic consistency modeling from the perspective of feature space alignment, contributing to a deeper understanding of the fundamental issues in multimodal fusion. By incorporating lightweight design principles, it also provides new insights into structural optimization of multimodal models.

To address the large parameter size and low inference efficiency of existing multimodal depression recognition models, this paper proposes a strategy that integrates lightweight network architectures with feature alignment mechanisms, reducing redundancy while maintaining competitive performance.

Lightweight models are more suitable for deployment on mobile devices and primary healthcare systems, facilitating low-cost and scalable early screening and assisted diagnosis of depression. This research therefore holds significant social and practical value.

2. Related Work

2.1. Deep Learning-Based Depression Recognition

With the advancement of deep learning technologies, neural network-based approaches have gradually replaced traditional machine learning methods for automatic depression recognition [4]. Early studies mainly employed Support Vector Machines (SVM) and Random Forests, relying heavily on handcrafted acoustic, textual, or facial action unit features. However, these approaches depend strongly on feature engineering and often suffer from limited generalization capability.

In recent years, Convolutional Neural Networks (CNNs),

Recurrent Neural Networks (RNNs), and their variants (such as LSTM and GRU) have been widely applied to depression recognition tasks. In the speech modality, CNNs are typically used to extract spectrogram features, while LSTMs model temporal dependencies. In the textual modality, pretrained language models such as BERT significantly enhance semantic representation of emotional content. In the video modality, 3D-CNNs or temporal convolutional networks are employed to capture dynamic facial expression patterns.

Furthermore, Transformer architectures have been introduced due to their strong global modeling capability, enabling cross-temporal and cross-modal interactions. However, large-scale deep models usually contain massive parameters and high computational complexity, which limits their practical deployment [5].

2.2. Multimodal Deep Fusion Methods

Multimodal depression recognition typically adopts three fusion strategies: early fusion, late fusion, and hybrid (intermediate) fusion [6].

Early fusion concatenates features from different modalities at the input stage, but it is sensitive to modality-specific noise. Late fusion independently models each modality and combines decisions at the output level. Although structurally simple, it fails to capture cross-modal interactions effectively.

Recently, intermediate fusion methods have become dominant [7]. Representative approaches include attention-based fusion networks, cross-modal Transformer architectures, and graph neural network-based frameworks. These methods model inter-modal dependencies and enable more effective information interaction. However, most fusion networks are structurally complex, parameter-intensive, and computationally expensive.

2.3. Cross-Modal Feature Alignment Methods

Due to heterogeneous statistical distributions and semantic spaces across modalities, direct fusion may lead to inconsistent feature representations. Therefore, cross-modal feature alignment has become a critical research topic in multimodal learning.

Mainstream approaches include:

Contrastive Learning-based Methods: These methods maximize similarity between positive cross-modal pairs while minimizing similarity between negative pairs, achieving semantic alignment across modalities.

Adversarial Learning-based Methods: Adversarial networks are employed to reduce distribution discrepancies between modality-specific features.

Shared Latent Space Mapping Methods: Projection networks map heterogeneous modality features into a unified semantic space.

Although these methods have achieved significant success in tasks such as image-text matching and audio-visual understanding, their application to depression recognition often involves complex architectures that are difficult to lightweight.

2.4. Lightweight Neural Network Design

To reduce model complexity, various lightweight strategies have been proposed [8], including:

(1) Depthwise separable convolution
(2) Model pruning
(3) Parameter quantization
(4) Knowledge distillation
(5) Lightweight Transformer variants (e.g., MobileViT, Tiny Transformers)

However, most lightweight research focuses on unimodal vision or speech tasks. Applications in multimodal depression recognition remain limited. In particular, achieving effective cross-modal alignment while maintaining lightweight design is still an open challenge.

In summary, although significant progress has been made in multimodal depression recognition, challenges remain regarding high model complexity and insufficient feature alignment. This paper addresses these issues by jointly optimizing lightweight architecture design and cross-modal alignment mechanisms, proposing an efficient yet effective multimodal depression recognition framework.

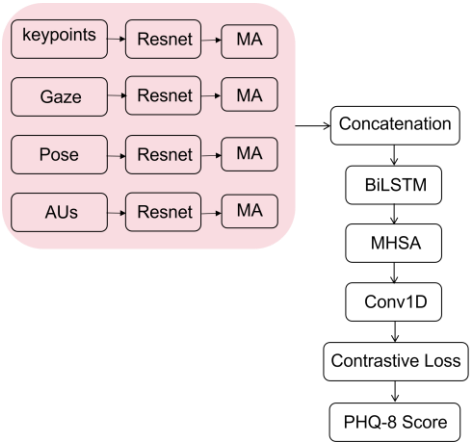

Figure 1. Overall architecture of the proposed lightweight multimodal feature alignment framework for depression severity estimation

3. Methodology

3.1. Cue-Aware Statistical Alignment

Heterogeneous visual cues originate from different physical measurement spaces and exhibit distinct statistical distributions. Direct cross-cue modeling may destabilize optimization due to scale inconsistency.

We therefore introduce cue-specific statistical alignment operators:

$$\hat{F}_i = Norm_i(F_i)$$

Each normalization module contains independent learnable parameters.

This mechanism:

(1) Preserves semantic direction
(2) Restructures statistical properties
(3) Establishes a unified scale space
(4) Introduces negligible overhead

It performs distributional restructuring rather than semantic projection.

3.2. Multi-Granularity Temporal Encoding

To capture depression-related dynamics at different temporal scales, we design a multi-granularity temporal encoding mechanism:

(1) Instance-level modeling
(2) Cue-level modeling

Given aligned representations:

$$H = BiLSTM(F)$$

The bidirectional LSTM captures long-term dependencies and provides contextualized representations.

This dual-level design enhances both micro-dynamic and macro-semantic modeling.

3.3. Gated Residual Attention Modulation

A gated residual attention mechanism is introduced after temporal encoding.

Attention weights are computed as:

$$A = Softmax\left(\frac{Q K^T}{\sqrt{d}}\right)$$

A sigmoid gating function produces modulation weights:

$$G = \sigma(A)$$

The final representation is:

$$\tilde{H} = H \odot G$$

Unlike standard attention, this design treats attention as a modulation mask rather than a direct replacement.

3.4. Cross-Cue Convolutional Aggregation

Cue representations are aggregated via 1D convolution:

$$Z = Conv1D\,(Concat\,(\tilde{H}_1, \ldots, \tilde{H}_K))$$

This operation models local cross-cue interactions more effectively than direct concatenation.

3.5. Representation Consistency Regularization

A consistency-based contrastive regularization is applied to enforce structural similarity across independently encoded representations.

This enhances intra-class compactness and inter-class separability.

4. Experiments

4.1. Dataset

We conduct experiments on the DAIC-WOZ dataset, a widely used benchmark for depression analysis. The dataset consists of clinical interviews designed to support the diagnosis of psychological distress conditions such as depression.

Following prior work, we formulate the task as a regression problem, aiming to predict the PHQ-8 score (ranging from 0 to 24), which measures depression severity.

4.2. Evaluation Metrics

We evaluate model performance using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which are widely adopted metrics for regression-based depression severity prediction. Lower values indicate better performance.

$$[MAE = \frac{1}{N}\sum_{i=1}^{N} |\,y_i - \hat{y}_i|$$

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

4.3. Implementation Details

We train the model using the Adam optimizer with an initial learning rate of 1e-4 and batch size of 32. The loss function is Mean Squared Error (MSE). Early stopping is applied based on validation performance. The official train/validation/test split of the DAIC-WOZ dataset is adopted. All experiments are repeated three times and the average results are reported.

4.4. Comparison with Baselines

Table 1. Performance comparison of different methods. Lower is better

Method	MAE	RMSE
RF [9]	5.88	7.13
Statistics + SVR [10]	4.91	5.98
ConvBiLSTM [11]	6.17	8.00
DepArt-Net [12]	4.61	5.78
Ours	4.44	5.77

We compare our method with representative traditional machine learning approaches (RF, Statistics + SVR) and deep learning-based temporal models (ConvBiLSTM, DepArt-Net).

As shown in Table 1, our method achieves an MAE of 4.44 and RMSE of 5.77.

5. Conclusion

In this study, we proposed a novel deep learning framework for depression detection. The proposed model is designed to capture complex feature interactions and temporal patterns that are highly relevant to depressive symptoms. By effectively modeling these characteristics, the framework enhances the representation capability of mental health-related signals.

Experimental results demonstrate that the proposed approach achieves competitive performance compared with several representative baseline methods. The findings suggest that the model is capable of extracting meaningful patterns associated with depressive states and provides stable predictive performance.

Importantly, the proposed framework is intended as a supportive tool rather than a replacement for clinical diagnosis. It may assist healthcare professionals in early screening and large-scale assessment scenarios, where efficient and automated analysis is required.

In future work, we plan to further improve model interpretability to better align with clinical understanding, explore multi-modal data integration, and validate the framework on larger and more diverse populations to enhance its generalizability.

Overall, this work contributes to the development of intelligent mental health assessment systems and provides a promising direction for depression detection research.

References

[1] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," Speech Communication, vol. 71, pp. 10–49, 2015.

[2] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in Proceedings of Interspeech, 2018, pp. 1716–1720.

[3] T. Baltruaitis, C. Ahuja, and L. philippe Morency, "Multimodal machine learning: A survey and taxonomy," IEEE

Transactions on Pattern Analysis and Machine Intelligence, vol. 41, pp. 423–443, 2017.

[4] J. Williamson, T. Quatieri, B. Helfer, R. Horwitz, B. Yu, and D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC 2013), 10 2013.

[5] N. Esmi, A. Shahbahrami, G. Gaydadjiev, and P. de Jonge, "Multimodal transformer for depression detection based on eeg and interview data," Biomedical Signal Processing and Control, vol. 113, p. 109039, 2026.

[6] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39–58, 2009.

[7] S. Patel, N. Shroff, and H. Shah, "Multimodal sentiment analysis using deep learning: A review," in Advancements in Smart Computing and Information Security, S. Rajagopal, K. Popat, D. Meva, and S. Bajeja, Eds.Cham: Springer Nature Switzerland, 2024, pp. 13–29.

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4510–4520.

[9] F. Ringeval, M. Pantic, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, and M. Schmitt, "Avec 2017: Real-life depression and affect recognition workshop and challenge," in Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge (AVEC 2017), 2017, pp. 3–9.

[10] S. Song, L. Shen, and M. Valstar, "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," in Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, ser. Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018. United States: Institute of Electrical and Electronics Engineers Inc., Jun. 2018, pp. 158–165, publisher Copyright: © 2018 IEEE.; 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018; Conference date: 15-05-2018 Through 19-05-2018.

[11] P.-C. Wei, K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, "Multi-modal depression estimation based on subattentional fusion," in ECCV Workshops, 2022.

[12] Z. Du, W. Li, D. Huang, and Y. Wang, "Encoding visual behaviors with attentive temporal convolution for depression prediction," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019, pp. 1–7.