Prior-Prompt-Based GCN for Depression Recognition Through Gait Observation

Chengju Zhou (cjzhou@scnu.edu.cn)

Yutao Xu (2024024547@m.scnu.edu.cn)

Yan Liang* (liangyan@m.scnu.edu.cn)
School of Artificial Intelligence, South China Normal University, Shishan Town, Nanhai District, Foshan, 528225, Guangdong Province, China

Abstract

In recent years, depression, as a prevalent mental health disorder has drawn increasing attention. With the advance of AI technology, automatic and objective diagnosis methods emerge by observing signals like electroencephalogram (EEG) signals, faces and behaviors. In the present paper, we propose gait analysis as a non-invasive method for depression detection. In this study, we propose a prior-prompt-based graph convolution network (PP-GCN) for depression recognition through gait that integrates skeleton and text modalities. Different from the conventional single-modal methods in the present study, we utilize prior knowledge and angle features. We innovatively introduce Generative Action Prompt (GAP), leveraging a pre-trained large language model to generate motion descriptions for different body parts, thereby providing prior knowledge for depression recognition. Additionally, considering the subtle gait feature variations in individuals with depression, we further incorporate a joint-angle-based representation strategy to capture fine-grained variations in movements. Experimental results demonstrate that the proposed model outperforms existing skeleton-based approaches on a large-scale dataset which contains over 25,000 gait sequences from nearly 300 volunteers named D-Gait, achieving excellent performance.

Keywords: depression; gait recognition; multi-modal; prompt learning

Introduction

Depression is one of the most common mental disorders, affecting over 300 million people worldwide (Friedrich, 2017). It not only significantly impacts patients' physical and mental health, but it can also lead to suicidal tendency (AlAzzam et al., 2021). As the number of depression patients and the risk of suicide continue to rise, there is an increasing demand for early detection of depression. Traditional diagnostic methods typically require patients to visit a hospital for professional psychological evaluation, where doctors make comprehensive judgments based on medical history, physical conditions, and psychological scales (Faust & Ziskin, 1988).

However, this diagnostic model faces two main challenges:

(1) The risk for inaccurate diagnosis due to patients intentionally withholding or misrepresenting information in questionnaire-based assessments, which can lead to flawed analysis and affect treatment decisions.

(2) Traditional methods require substantial human, material, and time resources in large-scale screening scenarios. Therefore, developing fast and efficient depression detection methods is of great significance.

Currently, commonly used objective detection methods mainly rely on EEG, facial expressions, eye movement data and others (Ay et al., 2019; Hu & Tao & Yang, 2023; Alghowinem et al., 2013). Although these methods have already achieved relatively high accuracy, they often require controlled experimental environments and may cause discomfort or resistance in patients, making them unsuitable for unobtrusive detection in public settings. Against this backdrop, gait information has gained increasing attention from researchers due to its ability to be collected remotely and non-intrusively. Existing studies have shown that the motor behaviors of individuals, particularly gait attributes, are significantly associated with depression. For example, individuals with depression often exhibited shorter stride length, slower walking speed, a forward-leaning posture, and reduced range of motion (Michalak et al., 2009). Current gait-based depression detection methods mainly fall into two categories: machine learning and deep learning.

Traditional machine learning methods typically rely on algorithms like support vector machines, logistic regression, and random forests to model and classify hand-crafted gait features (Peng & Hu & Dang, 2019; Wang et al., 2021; Fang et al., 2019). Although these methods have been successful in some cases, their main issue lies in the need for complex feature engineering and model tuning when processing raw data, making them less effective in handling more complex natural signal processing tasks.

In recent years, deep learning methods for depression detection have become the focus of research. These methods can be divided into two types: one is skeleton-based research, which uses the coordinates of skeleton joints for depression assessment (Wang et al., 2020). However, the challenge with this approach is that using only raw skeleton coordinates provides limited information, which may not be sufficient to capture the subtle gait variations induced by depression. Without additional contextual information, such as textual cues, the model may struggle to accurately identify these subtle differences in movement. The other type of method involves multi-modal fusion, where multiple modalities, such

* Corresponding author

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

as skeleton and silhouette data, are used simultaneously to enhance the model recognition and generalization ability through the complementary nature of different modalities (Shao et al., 2021). However, this approach faces the issue that silhouette data acquisition can be influenced by lighting conditions, and the fusion of modalities does not fully utilize some existing prior knowledge.

In recent years, multi-modal fusion techniques involving both images and text have gained significant attention. For example, the CLIP model constructs cross-modal representations by leveraging large-scale image-text data alignment and visual-language pre-training (Radford et al., 2021). In skeleton-related tasks (such as action recognition and gait analysis), researchers have also begun to introduce textual information to enhance model performance by utilizing prior knowledge (Xiang et al., 2023; Yang et al., 2024). This multi-modal fusion technology has significantly improved model performance in specific tasks.

Inspired by the above studies, we propose a prior-prompt-based graph convolution network (PP-GCN) for depression detection through gait that integrates skeleton and text modalities to enhance model performance. Specifically, we innovatively introduce a generative action description prompt (GAP), which leverages a pre-trained large language model to generate movement descriptions for body parts, providing prior knowledge for depression detection. Moreover, existing studies have shown that depression patients exhibit subtle differences in gait, such as slower walking, reduced vertical head movement, and sluggish arm swings (Michalak et al., 2009; Radovanović et al., 2014). These fine-grained gait features may be difficult to capture accurately when only using skeleton joint coordinates. To address these subtle differences in depression gait features, we go further to introduce an innovative angle representation strategy based on joints to capture fine-grained changes in movement.

We evaluate the proposed model PP-GCN on the D-Gait dataset (Liu et al., 2024). Experimental results show that this model outperforms existing skeleton-based methods and achieves excellent performance. Ablation studies further verify the significant improvement in model performance with the inclusion of prior knowledge and angle information.

Overall, our contributions can be summarized as follows:

(1) For the first time in the field of depression detection, we propose a prior-prompt-based model PP-GCN that leverages prior knowledge from large language models, significantly improving detection accuracy.

(2) We introduce a joint-angle-based feature extraction strategy to effectively capture fine-grained changes in movement for depression patients gait features.

(3) We validate the effectiveness of the model on the D-Gait dataset, significantly outperforming existing skeleton-based methods and achieving excellent performance.

Methodology

Model Architecture

An overview of our model is shown in Figure 1. In PP-GCN, skeleton data is processed through the Angle Feature Module (AFM) to extract angle features. These angle features, along with joint features, are then passed through a 1x1 convolution layer for dimensionality reduction before being input into the skeleton encoder. The skeleton encoder calculates the cross-entropy loss. Concurrently, the Generative Action Prompt (GAP) module generates text features for corresponding body parts. These text features are aligned with their corresponding skeleton features to compute the skeleton-text contrastive loss. The model is trained by combining the cross-entropy loss and the skeleton-text contrastive loss into an overall loss function.

Skeleton Encoder

In the field of skeleton-based action recognition, Graph Convolutional Networks (GCN) have been widely used due to their ability to closely match the structural characteristics of skeleton data. Among them, ST-GCN innovatively modeled human joints in both temporal and spatial dimensions (Yan & Xiong & Lin, 2018). However, the graph convolution operation in ST-GCN is based on a predefined skeleton topology, lacking the ability to dynamically adapt to different action features, and it performs relatively weakly in modeling the dependencies between channels.

To address the limitations of ST-GCN, CTR-GCN introduced a topology optimization mechanism based on the channel dimension, which can more effectively capture the dynamic correlations between keypoints. Moreover, the topology of CTR-GCN is not fixed but can adaptively learn based on the input data, thereby dynamically adjusting to accommodate different action patterns (Chen et al., 2021). Based on these improvements, CTR-GCN has been widely applied in skeleton-based action recognition tasks and has achieved significant results. Therefore, we chose CTR-GCN as the skeleton encoder in our framework. It should be noted that CTR-GCN was originally designed to handle 3D skeleton data, while our data only contains 2D information. To this end, we made appropriate modifications to CTR-GCN to enable it to effectively process 2D skeleton data.

The human skeleton can be represented as a graph, where joints serve as the vertices and bones as the edges. This graph can be denoted as $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}, \ldots\ldots, \boldsymbol{v_N}\}$ is the set of $\boldsymbol{N}$ vertices, $\boldsymbol{E}$ is the set of edges, typically represented in the form of an adjacency matrix $\boldsymbol{A} \in \boldsymbol{R^{N \times N}}$, with the element of the matrix $\boldsymbol{a_{i,j}}$ indicating the association strength between vertices $\boldsymbol{v_i}$ and $\boldsymbol{v_j}$ (Yan & Xiong & Lin, 2018). We denote the skeleton feature at the $\boldsymbol{l}$ layer of the skeleton encoder as $\boldsymbol{X^{(l)}}$, the skeleton feature at the $\boldsymbol{l+1}$ layer of the skeleton encoder as $\boldsymbol{X^{(l+1)}}$.

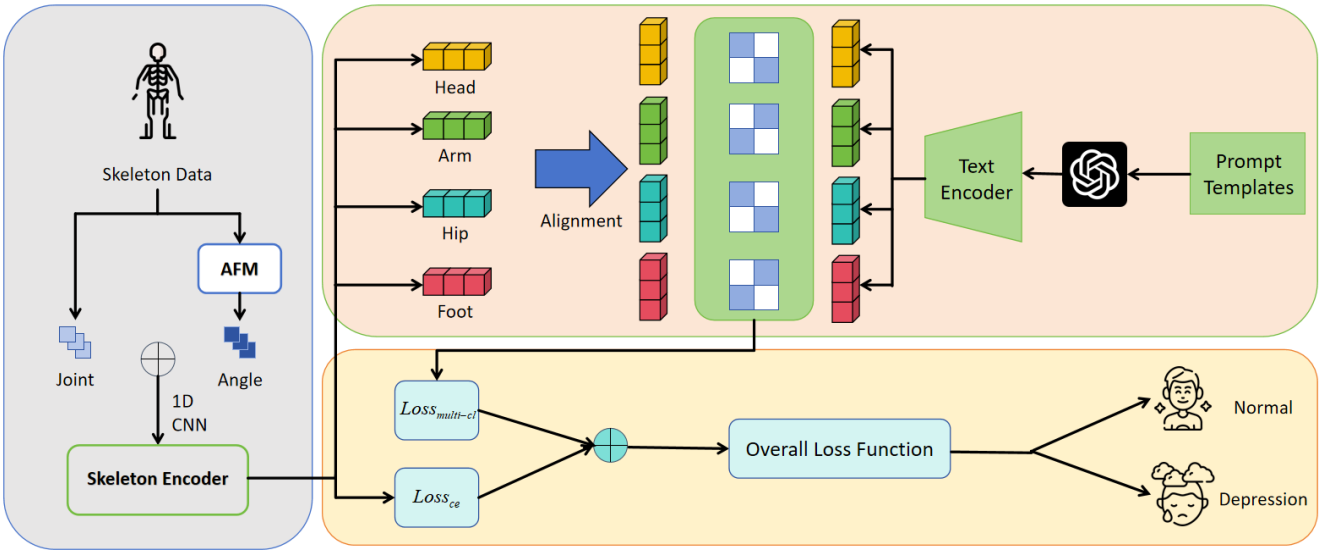

Figure 1: Overview of PP-GCN Network Based on the GAP and AFM.

The graph convolution process can be expressed by the formula:

$$X^{(l+1)} = \sigma(\Lambda^{\frac{1}{2}} A \Lambda^{-\frac{1}{2}} W^{(l)} X^{(l)}) \quad (1)$$

where Λ is a degree matrix, $W^{(l)}$ are the learnable parameters at the l layer, and σ is the activation function.

Generated Prompt Text Module

In recent years, with the rapid development of large language models, researchers have begun to combine them with the field of computer vision, promoting the progress of multi-modal learning. Among them, the CLIP model maps images and text to the same space, using contrastive learning to maximize the similarity of matching text-image pairs while minimizing the similarity of non-matching text-image pairs (Radford et al., 2021). The text information in the CLIP model is mainly obtained through preset prompt templates, such as "a photo of a {object}". In addition, prompt learning methods are also constantly evolving, with researchers introducing learnable prompt vectors to dynamically optimize prompt content in templates to meet the needs of different tasks (Zhou et al., 2022). However, the interpretability of current learnable prompt vectors remains an unresolved challenge.

For skeleton-based gait recognition, especially the fine-grained analysis of gait differences between individuals with depression and normal individuals, the human skeleton can be divided into different parts for in-depth study, and a large amount of prior knowledge has been accumulated in gait analysis. Based on this, we introduce the Generative Action Prompt (GAP). By utilizing GPT-4 as a knowledge engine, we designed prompt templates according with the differences in the characteristics of various body parts during the walking process of individuals with depression and normal individuals, thereby generating detailed descriptions of each body part during walking. This method helps to learn more refined skeleton representations, enable the model to more accurately capture the subtle differences in walking actions between individuals with depression and normal individuals, thereby improving the accuracy of the model.

(1) Multi-part Description Generation:

The text input plays a crucial role in the model's performance. To this end, we have experimented with various methods of generating text, such as the design approach of combining label names with prefixes and suffixes similar to the CLIP model (Radford et al., 2021), or manually defined text templates. However, the ablation experiment results (Xiang et al., 2023) in the GAP module indicate that the text information generated by pre-trained large language models can more efficiently integrate prior knowledge, providing the model with richer and more comprehensive information.

Based on this finding, we designed prompt templates for depression recognition, enabling the large language model (GPT-4) to generate action descriptions of various body parts during the walking process of individuals with depression and normal individuals. These text descriptions are then aligned with the skeleton features extracted by the skeleton encoder. Through contrastive learning, we maximize the similarity of matching skeleton-text pairs while minimizing the similarity of non-matching skeleton-text pairs. In this way, we effectively amplify the potential differences in movement patterns between individuals with depression and normal individuals, helping the model to more accurately learn key features. Table 1 shows the prompt templates we designed and the example descriptions they generated.

(2) Skeleton-Text Multi-part Contrastive Learning:

Traditional skeleton classification methods typically model skeleton features as a whole (holistic feature) and use one-hot labels for supervision. Although this approach is simple and intuitive, it has limitations in capturing fine-grained semantic information, especially in modeling local subtle changes in action features. In contrast, natural language supervision

Table 1: Designed Prompt Templates and Generated Descriptions.

Prompt Template
Describing following body parts actions when "a person suffering from depression walking": head, hand, arm, hip, leg, foot.
Generated Descriptions
head tilted slightly downward; hand may remain idle or loosely held, sometimes clenched subtly if tension or anxiety accompanies the depressive state; arm may swing less actively than usual during walking; hip movement may be reduced, resulting in a less dynamic or stiff gait; leg might lack a strong forward drive, emphasizing fatigue or heaviness; feet might drag slightly with each step.

through skeleton-text contrastive learning not only provides richer fine-grained semantic information for skeleton features but also significantly enhances the model's generalization and interpretability.

Therefore, we employ skeleton-text contrastive learning to optimize the skeleton encoder and text encoder in the model. The predicted similarity score between skeleton and text can be calculated using the following formula:

$$s_{i,j} = \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{N} exp(sim(z_i, z_k)/\tau)} \quad (2)$$

where z_i and z_j are the features output by the skeleton and text encoders respectively, τ is the temperature parameter, and Nrepresents the number of samples in the batch. This formula calculates the predicted similarity score between skeleton and text features.

After obtaining the similarity scores, we use KL divergence to calculate the loss function for skeleton-text contrastive learning:

$$Loss_{cl} = \frac{1}{2}[KL(s_{i,j}, t_{i,j}) + KL(s_{j,i}, t_{j,i})] \quad (3)$$

where $t_{i,j}$ and $t_{j,i}$ represent the true and predicted similarity scores between skeleton and text respectively. By training and optimizing over the entire dataset, the model can more effectively capture the semantic associations between skeleton and text.

Additionally, considering that the human skeleton is composed of multiple joint areas, modeling from multiple areas can better utilize prior knowledge. We divide the human skeleton into four parts: head, arms, hips, and legs, as shown in Figure 2. Correspondingly, we divide the output features of the skeleton encoder into four parts and match them with the text information generated by the large language model to calculate the skeleton-text contrast loss for each part separately. Finally, the multi-part skeleton-text contrast loss can be expressed by the following formula:

$$Loss_{multi-cl} = \frac{1}{M}\sum_{m=1}^{M} Loss_{cl}^{(m)} \quad (4)$$

where M represents the number of divide parts.

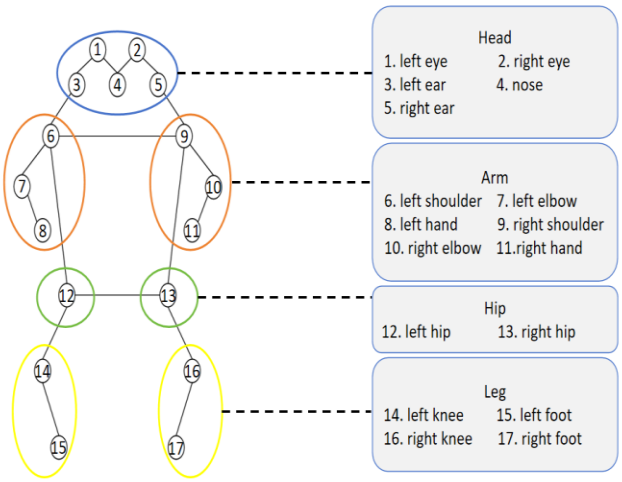

Figure 2: Four part partition strategies.

Angle Feature Module

Existing studies have shown that the gait of individuals with depression typically exhibits some subtle differences (Michalak et al., 2009; Radovanović et al., 2014). However, when using only skeleton joint data as model input, these fine-grained features are often difficult to accurately capture. Therefore, we introduce a method for extracting angle features based on joint points. By using angle features as additional input, the model can better capture the subtle differences in gait, thereby improving the recognition ability of gait features in individuals with depression.

To more accurately capture the changes in the angle features of joints during movement, we selected four groups of joints—hands, elbows, knees, and legs—as the anchor nodes for calculating angles, as shown in Figure 3. When calculating specifically, taking a group of joints as the anchor point, we sequentially traverse all other nodes in the skeleton graph and calculate the angle features of the target node using the following formula (Qin et al., 2022):

$$Angle_t = \begin{cases} 1 - \cos\theta = 1 - \dfrac{\vec{b}_{t,a_1} \cdot \vec{b}_{t,a_2}}{|\vec{b}_{t,a_1}||\vec{b}_{t,a_2}|}, t \neq a_1, t \neq a_2 \\ 0, \qquad\qquad\qquad\qquad t = a_1, t = a_2 \end{cases} \quad (5)$$

where $\vec{b}_{t,a_1} = (x_{a_1} - x_t, y_{a_2} - y_t)$ represent the vectors between the target node and the anchor points, and the length of the vector can be calculated through its coordinates.

After incorporating the angle features into the model, the input data is expanded from the original two-dimensional coordinates to three-channel information. Then, since the modified skeleton encoder is mainly suitable for processing two-dimensional information, we added a convolutional layer to compress the three-channel input into two channels, thereby effectively integrating the angle features and optimizing the model input structure.

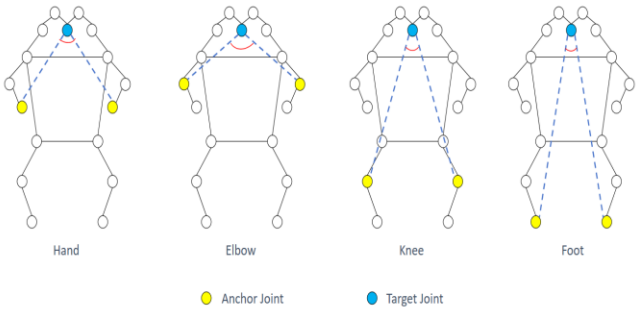

Figure 3: Joint Groups Used for Angle Calculation.

Training And Prediction

The training of the skeleton encoder is achieved through the cross-entropy loss function, which aims to minimize the difference between the predicted distribution and the true labels. The form of the loss function is:

$$Loss_{ce} = -\sum_{i=1}^{N} y_i \, log(\hat{y}_i) \tag{6}$$

where y^i is the true label of the sample i, $\hat{y}_i$ is the predicted probability distribution output by the skeleton encoder, and N is the total number of samples.

The total loss function used for training the model can be expressed as:

$$Loss_{total} = Loss_{ce} + \lambda Loss_{multi-cl} \tag{7}$$

where λ is a hyperparameter used to balance the multi-part contrastive loss and the cross-entropy loss.

Experiments

Datasets And Settings

We evaluated the performance of our model on the D-Gait dataset (Liu et al., 2024). This dataset contains over 25,000 gait sequences from nearly 300 volunteers, captured from 16 different viewpoints and under 3 different clothing conditions. The data includes both skeleton and silhouette modalities. The depression risk of the volunteers was assessed using three standard psychological scales: PHQ-9 (Kroenke & Spitzer & Williams, 2021), SDS (Zung, 1965), and GAD-7 (Spitzer et al., 2006). Based on the scores from these scales, the data was comprehensively labeled as either "depressed" or "normal".

Model configurations: The skeleton encoder accepts 2D coordinates ($input\ channels = 2$) while using CLIP's pre-trained text encoder. The class weight of cross-entropy loss is set to a ratio of 1:2 for normal and depressed. The hyperparameter λ used to balance the contrastive loss and cross-entropy loss is set to 1.0. Training employs SGD with OneCycleLR ($peak\ lr = 0.01$), batch sizes 64/128 (train/test) for 100 epochs. Angle features from hands, elbows, knees and feet are decision-fused with weights [0.8:1:0.2:0.2]. Our model is implemented in Pytorch on RTX 4090 GPU.

Evaluation Metrics

While individual precision and recall metrics exhibit variability across ablation studies, we prioritize the F1-score as the primary evaluation criterion due to its ability to balance both precision and recall. This is particularly crucial in depression recognition tasks, where clinical applicability demands a balance between sensitivity (recall) and diagnostic accuracy (precision). The F1-score effectively captures the trade-offs between these objectives.

Comparisons with Other Methods

The methods used for comparison in the experiments are as follows.

(1) ST-GCN (Yan & Xiong & Lin, 2018): An action recognition method that models both spatial and temporal information of the skeleton simultaneously.

(2) CTR-GCN (Chen et al., 2021): Based on ST-GCN, it proposes a topology optimization mechanism on the channel dimension, which can dynamically learn the topological structure of different actions.

We evaluated these methods on the D-Gait dataset and compared them with our proposed method, which is based on CTR-GCN combined with the GAP and AFM modules.

Table 2: Performance Comparison on D-Gait Dataset.

Models	Precision (%)	Recall (%)	F1-score (%)
ST-GCN	46.01	69.40	55.34
CTR-GCN	46.68	72.25	56.71
PP-GCN	**47.74**	**73.59**	**57.91**
△SOTA	**1.06** ↑	**1.34** ↑	**1.20** ↑

As shown in Table 2, our method achieved significant improvements in precision, recall and F1 score, realizing the best performance to date. The experimental results on the D-Gait dataset are as follows: precision 47.74%, recall 73.59%, and F1 score 57.91%. This performance is attributed to our full exploration and utilization of existing prior knowledge, integrating it into the model in the form of text information, and introducing angle features to capture the subtle differences in gait details, which significantly enhanced the model's performance in the depression recognition task.

Ablation Study

We validated the contributions of each module to the model performance through ablation experiments, with all experiments conducted on the D-Gait dataset. We selected the F1-score as the final evaluation metric because it provides a balanced measure by combining both Precision and Recall.

(1) Performance with Various CLIP Pre-trained Text Encoders:

When we use different CLIP pre-trained text encoders in the model, the model's performance varies differently.

Table 3: Performance with Various CLIP Pre-trained Text Encoders (without AFM).

TextEncoder	Precision (%)	Recall (%)	F1-score (%)
ViT-B/32	46.13	74.62	57.11
ViT-B/16	45.19	75.96	56.77
ViT-L/14	49.87	67.60	**57.41**
RN50x64	45.63	73.71	56.47

As shown in Table 3, experimental results indicate that choosing ViT-L/14 as the text encoder can achieve a good balance between efficiency and accuracy, so we ultimately use it as the model's text encoder.

(2) Performance with Different Angle Feature Combinations:

We experimented with the impact of incorporating angle features from different body parts on model performance, as well as the results of fusing multi-part angle features at the feature and decision levels.

Table 4: Performance with Different Angle Feature Combinations.

AngleFeature	Precision (%)	Recall (%)	F1-score (%)
Hand	45.69	76.31	57.16
Elbow	48.28	69.73	57.06
Knee	47.61	71.34	57.11
Foot	45.57	74.18	56.46
Feature Fusion	47.65	68.62	56.24
Decision Fusion	47.74	73.59	**57.91**

As shown in Table 4, the experimental results indicate that when only the angle features of a certain part are added, the model performance actually decreases. However, when the angle features of the hands, elbows, knees, and feet are fused at the decision level according to the weight ratio [0.8:1:0.2:0.2], the model performance reaches the best and has a significant improvement. This shows that considering the angle features of multiple parts comprehensively can better optimize the model effect, and the angle changes of the hands and elbows are more important for depression detection.

Visualization Analysis

To verify the effectiveness of the GAP and AFM, we calculate and visualize the confusion matrix for our model.

As shown in Figure 4, based on the visualization of the confusion matrix. We found that the GAP module, by integrating textual information, significantly enhanced the model ability to capture subtle gait differences, thereby improving the detection of true positives (TP) and true negatives (TN). Meanwhile, the AFM module, by extracting joint angle features, further strengthened the model ability to distinguish the subtle gait differences between depressed individuals and normal ones. However, the false positive (FP) rate of 37.81% and the false negative (FN) rate of 26.41% indicate that the model still needs to improve its ability to capture subtle gait differences and reduce misclassifications.

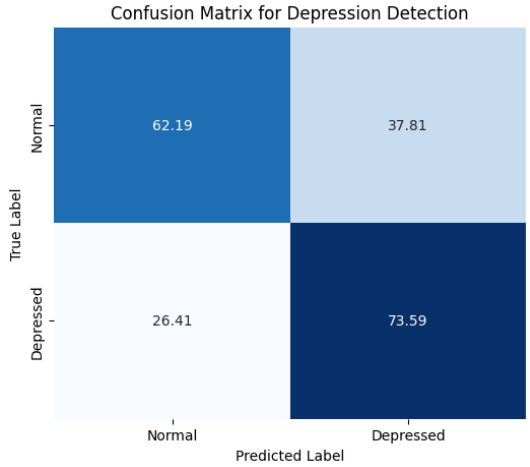

Figure 4: Confusion Matrix for Depression Detection.

Conclusion

This paper proposes a prior-prompt-based graph convolution network (PP-GCN) for depression recognition through gait, which integrates a Generative Action Prompt (GAP) and an Angle Feature Module (AFM). The GAP leverages a pre-trained large language model to generate descriptions of movements for different body parts, providing rich prior knowledge for depression recognition and addressing the insufficient use of prior knowledge in traditional methods. The AFM captures subtle differences in gait by extracting joint angle features, thereby enhancing the model's ability to identify gait characteristics of individuals with depression. Ablation experiments on the D-Gait dataset demonstrate that the introduction of the GAP and AFM modules significantly improves model performance, with the best results achieved when fusing angle features from multiple body parts. These innovations not only resolve the shortcomings of traditional gait recognition methods in terms of generalization and classification performance but also offer a new perspective for early depression detection and point the way for future research in this area.

Acknowledgments

This work was partially supported by the Guangdong Basic and Applied Basic Research Foundation under grant 2025A1515011526.

References

Friedrich, M. J. (2017). Depression is the leading cause of disability around the world. *Jama*, *317*(15), 1517-1517.

AlAzzam, M., Abuhammad, S., Tawalbeh, L., & Dalky, H. (2021). Prevalence and correlates of depression, anxiety, and suicidality among high school students: a national study. *Journal of psychosocial nursing and mental health services*, *59*(8), 43-51.

Faust, D., & Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science*, *241*(4861), 31-35.

Ay, B., Yildirim, O., Talo, M., Baloglu, U. B., Aydin, G., Puthankattil, S. D., & Acharya, U. R. (2019). Automated

depression detection using deep representation and sequence learning with EEG signals. *Journal of medical systems*, *43*, 1-12.

Hu, B., Tao, Y., & Yang, M. (2023). Detecting depression based on facial cues elicited by emotional stimuli in video. *Computers in Biology and Medicine*, *165*, 107457.

Alghowinem, S., Goecke, R., Wagner, M., Parker, G., & Breakspear, M. (2013, September). Eye movement analysis for depression detection. In *2013 IEEE International Conference on Image Processing* (pp. 4220-4224). IEEE.

Michalak, J., Troje, N. F., Fischer, J., Vollmar, P., Heidenreich, T., & Schulte, D. (2009). Embodiment of sadness and depression—gait patterns associated with dysphoric mood. *Psychosomatic medicine*, *71*(5), 580-587.

Peng, Z., Hu, Q., & Dang, J. (2019). Multi-kernel SVM based depression recognition using social media data. *International Journal of Machine Learning and Cybernetics*, *10*, 43-57.

Wang, Y., Wang, J., Liu, X., & Zhu, T. (2021). Detecting depression through gait data: examining the contribution of gait features in recognizing depression. *Frontiers in psychiatry*, *12*, 661213.

Fang, J., Wang, T., Li, C., Hu, X., Ngai, E., Seet, B. C., ... & Jiang, X. (2019). Depression prevalence in postgraduate students and its association with gait abnormality. *IEEE Access*, *7*, 174425-174437.

Wang, T., Li, C., Wu, C., Zhao, C., Sun, J., Peng, H., ... & Hu, B. (2020). A gait assessment framework for depression detection using kinect sensors. *IEEE Sensors Journal*, *21*(3), 3260-3270.

Shao, W., You, Z., Liang, L., Hu, X., Li, C., Wang, W., & Hu, B. (2021). A multi-modal gait analysis-based detection system of the risk of depression. *IEEE Journal of Biomedical and Health Informatics*, *26*(10), 4859-4868.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Xiang, W., Li, C., Zhou, Y., Wang, B., & Zhang, L. (2023). Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10276-10285).

Yang, Z., Wu, D., Wu, C., Lin, Z., Gu, J., & Wang, W. (2024). A Pedestrian is Worth One Prompt: Towards Language Guidance Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 17343-17353).

Michalak, J., Troje, N. F., Fischer, J., Vollmar, P., Heidenreich, T., & Schulte, D. (2009). Embodiment of sadness and depression—gait patterns associated with dysphoric mood. *Psychosomatic medicine*, *71*(5), 580-587.

Radovanović, S., Jovičić, M., Marić, N. P., & Kostić, V. (2014). Gait characteristics in patients with major depression performing cognitive and motor tasks while walking. *Psychiatry research*, *217*(1-2), 39-46.

Qin, Z., Liu, Y., Ji, P., Kim, D., Wang, L., McKay, R. I., ... & Gedeon, T. (2022). Fusing higher-order features in graph neural networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(4), 4783-4797.

Liu, X., Li, Q., Hou, S., Ren, M., Hu, X., & Huang, Y. (2024). Depression risk recognition based on gait: A benchmark. *Neurocomputing*, 128045.

Yan, S., Xiong, Y., & Lin, D. (2018, April). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., & Hu, W. (2021). Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13359-13368).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, *130*(9), 2337-2348.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), 606-613.

Zung, W. W. (1965). A self-rating depression scale. *Archives of general psychiatry*, *12*(1), 63-70.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, *166*(10), 1092-1097.