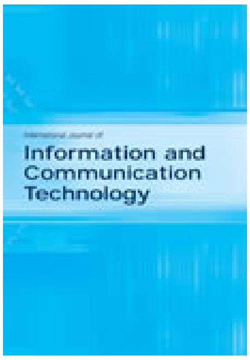


International Journal of
Information and
Communication
Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Enhancing accuracy of pragmatic ability tests through multi-feature fusion based on graph neural networks

Teng Xie, Dingyu Liu, Wei Zhou

DOI: 10.1504/IJICT.2025.10075259

Article History:

Received:	22 October 2025
Last revised:	15 November 2025
Accepted:	15 November 2025
Published online:	12 January 2026

Enhancing accuracy of pragmatic ability tests through multi-feature fusion based on graph neural networks

Teng Xie and Dingyu Liu

College of Teacher Education,
Longyan University,
Fujian 364000, China
Email: 82012005@lyun.edu.cn
Email: ding1232025@126.com

Wei Zhou*

College of Foreign Languages,
Longyan University,
Fujian 364000, China
Email: clarezhouwei@lyun.edu.cn
*Corresponding author

Abstract: Pragmatic ability assessment holds significant importance in language teaching and related fields, yet existing methods fail to capture and utilise the characteristics and information across different modalities. To address this, this paper optimises graph neural networks through multi-stage adaptive fusion. By decomposing the graph neural network into a multi-stage training format, higher-order features of graph data are progressively integrated into shallow models across multiple stages, thereby training a more robust shallow model. Subsequently, a pragmatic competence prediction model based on an improved graph neural network and multi-feature fusion is proposed. First, modal information is progressively integrated to ensure comprehensive fusion. Then, long-range pragmatic information is captured and incorporated into sentence-level information, enabling the model to better understand global features. Experimental results demonstrate that the proposed model achieves at least a 3.46% improvement in pragmatic competence test accuracy, facilitating more precise assessment of pragmatic competence levels.

Keywords: pragmatic competency assessment; graph neural network; GNN; multimodal feature; multi-stage optimisation; adaptive fusion.

Reference to this paper should be made as follows: Xie, T., Liu, D. and Zhou, W. (2025) 'Enhancing accuracy of pragmatic ability tests through multi-feature fusion based on graph neural networks', *Int. J. Information and Communication Technology*, Vol. 26, No. 50, pp.133–149.

Biographical notes: Teng Xie is an Associate Professor in the College of Teacher Education at Longyan University, China. He received a Master's degree in Basic Psychology from Fujian Normal University. His research areas include child cognition and development.

Dingyu Liu is an Associate Professor in the College of Teacher Education at Longyan University, China. She received her Master's degree in Early Childhood Education from Fujian Normal University. Her research areas include basic education and teacher professional development.

Wei Zhou is a Teacher in the College of Teacher Education at Longyan University, China. She received her Master's degree (2014) and PhD (2018) from Soongsil University in South Korea. She researches areas include machine learning, heritage language maintenance and shift.

1 Introduction

Pragmatic competency, as a core component of linguistic ability, encompasses an individual's capacity to understand and use language effectively for communication in specific contexts. It involves not only the mastery of grammatical rules but also emphasises precise comprehension of the underlying social culture and communicative intentions (Prasatyo et al., 2023). In the era of deep integration between globalisation and informatisation, good pragmatic competency is essential for cross-cultural communication. However, traditional methods of pragmatic competency assessment face numerous challenges (Kentmen et al., 2023). On one hand, existing assessments often focus on language forms and simple semantics, making it difficult to comprehensively and thoroughly evaluate an individual's pragmatic performance in real complex contexts (Fathi et al., 2025). On the other hand, traditional assessment methods have limitations in dealing with the complexity and diversity of pragmatic data. Pragmatic phenomena are influenced by multiple factors, including linguistic and cultural background, the relationship between the interlocutors, and specific situations, which intertwine to make the features of pragmatic data highly nonlinear and interrelated (Alsuhaibani, 2022). Traditional statistical methods and machine learning models often struggle to fully extract potential information from this kind of complex data, affecting the accuracy and reliability of assessment results (Planques and Julián, 2018). Therefore, how to construct an efficient model to improve the accuracy of pragmatic competency assessment remains a crucial scientific issue.

Improving the accuracy of pragmatic competency assessment primarily relies on enhancing the accuracy of pragmatic competency prediction. Early researchers mainly adopted a technical approach based on manual feature engineering and classical machine learning algorithms. Typical implementation schemes included Naïve Bayes classifiers (Flores et al., 2014), support vector machines (Li et al., 2023), and decision tree models (Chowanda et al., 2021). This technical paradigm usually requires a complex text preprocessing workflow to construct traditional feature representation methods such as term frequency-inverse document frequency and bag-of-words models (Wahlster, 2023). Although these methods demonstrate good performance under limited datasets and low-dimensional feature spaces, their inherent architecture struggles to effectively capture long-distance contextual dependencies in pragmatic phenomena, and their capability to model deep semantic relationships is insufficient. Moreover, these systems are highly sensitive to language noise and semantic ambiguity, resulting in poor stability and generalisation performance in real-world application environments.

In recent years, deep learning algorithms have become the primary method for pragmatic competency prediction. Deep learning-based methods for predicting pragmatic competence exhibit characteristics such as automated feature engineering, robust context modelling, understanding of non-literal meanings, end-to-end learning, and multimodal fusion. Underpinning these features is the core principle of leveraging massive datasets through distributed representations, relying on context-aware architectures centred on the Transformer self-attention mechanism, and addressing complex pragmatic challenges via the pre-training-fine-tuning paradigm. Deep learning models typically consist of multi-layer neural networks, with the core focus on learning and extracting feature representations of various real-world entities from large datasets. These features can not only be used in various computational models but can also be directly processed and applied by computers. Due to the superior performance of deep learning models in natural language processing, numerous scholars have conducted research on pragmatic competency prediction using deep learning (Eragamreddy, 2025). Dai and Zhao (2022) utilised convolutional neural networks (CNN) for text processing and produced pragmatic competency predictions through a fully connected network. Kim et al. (2019) considered complex sentence structures and introduced tree-structured long short-term memory (LSTM) for pragmatic competency classification. To effectively model the representation of pragmatic documents, Ai et al. (2024) used CNN and LSTM to obtain sentence representations, and then employed gated recurrent neural networks (RNNs) to encode sentence semantics and their intrinsic relationships. Parola et al. (2021) developed a hierarchical attention network for pragmatic competency classification tasks, using an attention mechanism to help the network select important words and sentences. In addition to pragmatic competency prediction via text, researchers have also explored pragmatic competency prediction in psychiatry patients through images and speech. Sinclair et al. (2021) proposed a pragmatic competency prediction method based on foreground and background segmentation. This method is based on the YOLOv5 framework and introduces the ConvNeXt module and attention module for feature extraction and fusion, respectively, to improve pragmatic competency prediction accuracy. Zainal et al. (2024) applied the Transformer to pragmatic competency prediction, using a fused input of log-mel spectrograms and their first-order differential features, and utilised the Transformer to extract hierarchical speech representations, analysing the effects of changes in the number of attention heads and encoder layers on prediction accuracy.

Most of the aforementioned deep learning-based pragmatic competency prediction models are based on a single modality. Integrating features from different modalities, thereby achieving fusion of multimodal information, plays a crucial role in enhancing model training accuracy and compensating for the shortcomings of features from a single modality. The multimodal pragmatic competence assessment model's most significant advantage lies in its ability to transcend textual limitations, repositioning the evaluation anchor from language itself to the complete context in which language is used. This enables it to capture nuanced pragmatic subtleties that are often implicit rather than explicit, thereby far surpassing unimodal models in terms of assessment accuracy, depth, and human-centredness. Salamanti et al. (2023) incorporated both temporal and semantic consistency into the multimodal pragmatic competency prediction task, achieving pragmatic competency prediction through fine-grained temporal alignment and cross-modal semantic interaction. Chen (2023) proposed a multimodal deep regression Bayesian network (MMDRBN) to calculate the relationship between audio and visual modalities in the pragmatic competency prediction task and incorporated domain knowledge from the video; however, the prediction accuracy is not high. graph neural network (GNN) is based on graph-structured data and can effectively capture complex relationships between nodes and global information. Through information propagation and aggregation operations on the graph, GNN can learn feature representations of nodes at local and global levels, thus better understanding the intrinsic structure and semantic information of the data. He et al. (2022) proposed a pragmatic competency prediction model based on a heterogeneous graph, which is based on a heterogeneous GNN and performs unified modelling on multi-source information such as facial expressions, audio, and personality traits, thereby predicting pragmatic competency. Yan and Chen (2024) proposed a pragmatic competency prediction model based on a graph attention network and used a gated recurrent unit to capture complex interaction relationships between multimodal features, thereby improving the accuracy of pragmatic competency prediction.

Based on the analysis of current pragmatic competency prediction models, the existing methods have relatively simple modal fusion methods that cannot fully capture and utilise the characteristics and information of different modalities. Additionally, these methods focus more on capturing local contexts, especially when processing long conversations, often ignoring the integration of distant pragmatic features of speakers. To address these challenges, this paper proposes a pragmatic competency testing enhancement method based on GNN and multi-feature fusion. First, to address the issue where graph convolutional network (GCN) incurs excessive time and space consumption when the model depth increases, this paper proposes a GNN based on multi-stage adaptive fusion, called MSFGCN. MSFGCN divides the deep GNN model into a multi-stage training mode, with each stage containing several feature extraction layers. The main function of the deep learning module based on multi-stage training is to gradually integrate deep graph data information into a shallow model to train a more powerful shallow model. Then, a pragmatic competency prediction model based on MSFGCN and multi-modal feature fusion is proposed. This model consists of a multi-modal fusion module and a long-range sentiment fusion module. The multi-modal fusion module consists of three bi-directional fusion modules. Each bi-directional fusion module integrates multi-modal information from both forward and reverse directions, gradually fusing modal information to ensure thorough integration. The long-range feature fusion module first constructs sentence information from the pragmatic context, then captures long-range pragmatic information, and incorporates it into the sentence information, enabling the model to better understand global features. Finally, the Softmax function is used to obtain the pragmatic competency assessment results. Experimental results show that the accuracy and AUC of the proposed pragmatic competency assessment model significantly outperform those of baseline models, achieving precise pragmatic competency testing.

2 Relevant theory

2.1 Graph neural network

GNN has been widely applied in natural language processing tasks such as text sequence modelling and knowledge graph construction, thanks to its superior performance in handling unstructured data. Compared to traditional sequence-based deep learning methods like RNN and LSTM, GNN effectively captures the complex dependencies between text sequences by explicitly modelling nodes and their topological structures, thereby extracting deeper semantic feature representations. The information propagation process of GNN is shown in Figure 1. The core idea of GNN is to define nodes and their neighbourhoods, iteratively using the features of neighbouring nodes as the learning targets for each node. It utilises an update function to iteratively aggregate and update node states, thereby generating node representations that incorporate information from neighbouring nodes and graph topological structures. The formal representation of the above process is as follows (Zhou et al., 2022).

$$h_v^{(k)} = f\left(h_v^{(k-1)}, aggregate\left(\left\{h_u^{(k-1)} : u \in N(v)\right\}\right), X_v\right) \tag{1}$$

where $h_v^{(k)}$ represents the hidden state of node v at the k^{th} layer, $N(v)$ represents the neighbour set of node v, $aggregate(\cdot)$ is the aggregation function for neighbour information, and $f(\cdot)$ is the nonlinear update function. For the implementation of the node update function $f(\cdot)$ in GNN, researchers have proposed GCN and graph attentional neural networks (GAT) (Verma et al., 2023).

GCN learns node embedding representations by aggregating the features of neighbouring nodes, a process called convolution or neighbour aggregation. It can effectively utilise the features of nodes and their local information for prediction, thus demonstrating unique advantages in handling graph-structured data. However, as the number of layers in GCN increases, the embedding representations of nodes may become similar, causing the model to lose its discriminative capability. This issue is particularly pronounced in deep GCN. GAT uses a self-attention mechanism to enable each node in the graph to dynamically aggregate information from its neighbouring nodes based on their importance. This mechanism allows GAT to adaptively focus on more important neighbours for the current node, thus learning better node representations. Because GAT needs to compute attention weights for each node and all its neighbours, its computational cost is relatively high. Especially when dealing with large-scale graphs, the computational cost significantly increases. Therefore, in practical applications, appropriate GNN variants should be selected based on the specific scenario for research.

2.2 Feature fusion theory

Feature fusion is the process of merging multiple modal feature sets into a unified feature set. Through feature fusion, the complementarity of various features can be fully utilised, compensating for the limitations of single features and enhancing the model's ability to handle complex issues. Feature fusion is divided into early fusion, late fusion, and model-level fusion (Ma et al., 2016).

1 Early fusion. First, the original features need to be extracted from multiple modalities. Subsequently, fusion operations such as concatenation and summation are performed on these features to form the final feature representation. Simple concatenation or summation operations can easily introduce noise, which may have a negative impact on the final prediction.

2 Late fusion. It first involves separately training different modalities. Each modality is equipped with a dedicated classifier, enabling it to independently learn and extract modality-specific information. Fusion after each modality is processed independently may lead to the loss of some modality-specific information, thereby reducing the system's comprehensive understanding of the overall information.

3 Model-level fusion. Effective fusion of multiple modalities in dialogues is achieved through deep learning models, enabling mutual interaction and enhancement of information. As a comprehensive approach to multimodal fusion, model-level fusion enhances the performance of deep learning models in multimodal tasks by considering the correlations between models and promoting information interaction among different modalities (Li et al., 2025).

Early fusion directly concatenates or overlays multimodal features at the model input layer, forming a single feature vector for subsequent network processing. While simple and efficient, this approach suffers from modality heterogeneity and noise sensitivity issues. Late-stage fusion independently models each modality's features before integrating results at the decision layer. However, intermodal interactions occur only at the decision layer, failing to capture cross-modal correlations in shallow or intermediate layers. Model-level fusion enables dynamic intermodal interactions through mechanisms like attention layers or GNNs in intermediate layers, combining the advantages of early and late fusion. This paper employs model-level fusion for multimodal integration.

3 Optimisation of graph neural networks based on multi-stage adaptive fusion

To address the issue of excessive time and space consumption caused by deepening the number of layers in GCN, this paper proposes a GNN based on multi-stage adaptive fusion, called MSFGCN. MSFGCN divides GCN into a multi-stage training format, gradually integrating high-order features of graph data into a shallow model through multiple stages, thus training a more powerful shallow model. In addition, MSFGCN designs an adaptive fusion module based on an attention mechanism, which can adaptively train the fusion weights between deep features and shallow features.

Suppose the graph is defined in the form of $G = (V, E)$, where V is the set of nodes indexed starting from 1, and E is the set of edges between nodes in G. $N = |V|$ and $m = |E|$ represent the number of nodes and edges, respectively. In this section, we only consider an undirected and unweighted graph. The topological information of the entire graph is described by the adjacency matrix $A \in R^{n \times n}$, where $A_{(i,j)} = 1$ if there is an edge between node i and node j, otherwise $A_{(i,j)} = 0$. The diagonal matrix representing the node degrees is denoted as $D \in R^{n \times n}$, where $D_{(i,i)} = \sum_j A_{(i,j)}$. N_i indicates the set of adjacent nodes of node i. An attribute graph has a node feature matrix $X \in R^{n \times n}$, where each row $x_i \in R^d$ represents the feature vector of node i, and d is the dimension of node features.

Since each layer of the GCN model is equivalent to a low-pass filter, the graph data features after multiple low-pass filters become similar, that is, the over-smoothing problem. To alleviate this problem, this paper introduces the initial residual connections and identity mapping operations into the deep GCN to obtain deep features of the graph data, as shown below.

$$H^{(l+1)} = \sigma\left(\left(\left(1-\gamma^{(l)}\right)\hat{P}H^{(l)} + \gamma^{(l)}H^{(0)}\right)\left(1-\beta^{(l)}\right)I_n + \beta^{(l)}W^{(l)}\right)\right) \tag{2}$$

where $\hat{P} = \hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2} = \left(D+I_n\right)^{-1/2}\left(A+I_n\right)\left(D+I_n\right)^{-1/2}$, A is the adjacency matrix of the graph, and D is the degree matrix of the adjacency matrix. $W^{(l)}$ is the learnable parameters of the l^{th} level network. $H^{(l)}$ is the hidden features learned by the l^{th} level of the model. $\gamma$and $\beta$are hyperparameters that need to be manually set, and $\beta = log(\lambda/l + 1)$, $H^{(0)}$ are the initial features of the graph data, which are obtained from the original data through a fully connected network for dimensionality reduction. The dimensionality reduction operation is defined as follows.

$$H^{(0)} = MLP(X),\left(X \in R^{n\times d}, H^{(0)} \in R^{n\times f} f < d\right) \tag{3}$$

Since the GCN needs to compute the gradients of all nodes in the entire graph and also needs to store the embedding representations of all nodes, it will lead to a linear relationship between the time and space consumption required for model training and testing and the number of model layers. To address this, this paper proposes to split the deep feature extraction model into a multi-stage training approach and integrate the deep features of each stage into the shallow backbone model.

This paper divides the deep model into m stages, with each stage consisting of l/m graph convolution levels. Each stage can extract features of l/m levels. We then integrate the deep features extracted at each stage into the corresponding feature extraction layer of the shallow backbone model, as shown below.

$$H^{(m,l')} = \sigma\left(\left(\left(1-\gamma^{(m,l')}\right)\hat{P}H^{(m,l')} + \gamma^{(m,l')}H^{(0)}\right)\left((1-\beta^{(m,l')}\right)I_n + \beta^{(m,l')}W^{(m,l')}\right)\right) \tag{4}$$

$$F^{(m)} = fusion\left(\tilde{H}^{(m)}, H^{(m,l')}\right) \tag{5}$$

where $H^{(m,j)}$ is the l^{th} graph convolution layer of the m^{th} stage. $\tilde{H}^{(m)}$ is the shallow backbone feature of the m^{th} stage, *fusion*() is the fusion operation, and $F^{(m)}$ represents the feature after the fusion in the m^{th} stage.

After completing the deep feature extraction at each stage, we obtain two groups of features, which are the deep data features $H^{(l)}$ and the shallow backbone features $\tilde{H}^{(l)}$. The attention mechanism can be used to adaptively learn their fusion weights.

$$\left(\alpha_d, \alpha_s\right) = att\left(H^{(l)}, H^{(l)}\right) \tag{6}$$

where α_d and α_s are the learned fusion weights of the deep features $H^{(l)}$ and the shallow backbone features $\tilde{H}^{(l)}$, respectively. *att*() is the attention mechanism.

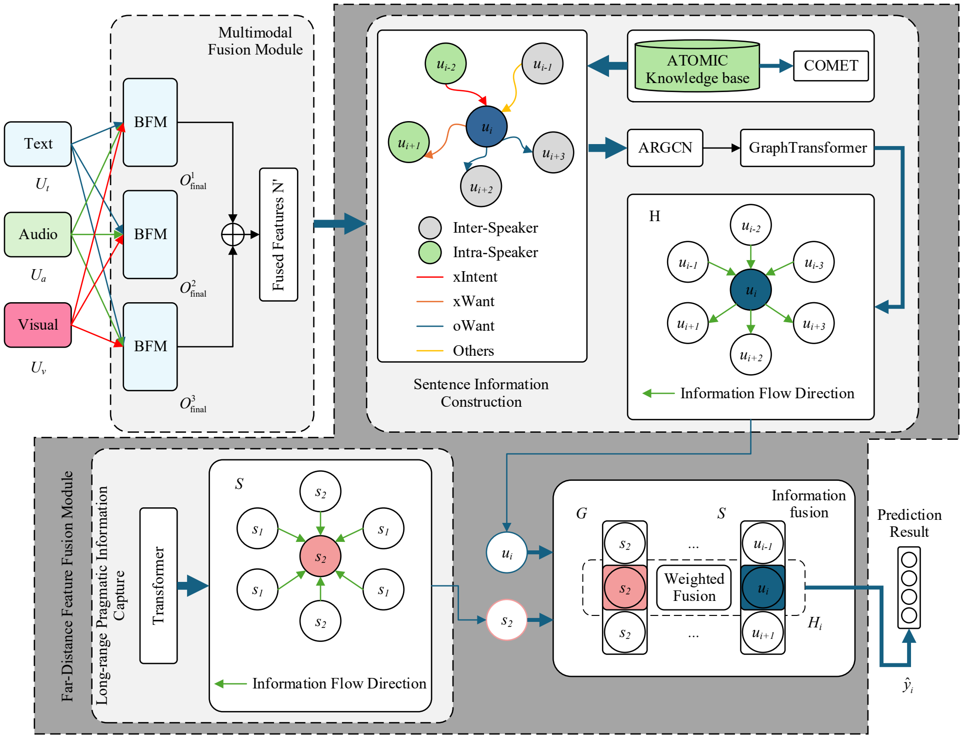
Based on the above operations, MSFGCN adaptively learns the fusion weights of deep features and shallow backbone features, and features with higher importance have larger fusion weights, which can better fuse the two sets of features. In MSFGCN, the deep model and the shallow backbone model are trained in parallel during training. During testing, this paper splits the deep sub-model from the backbone model and uses only the backbone model for testing, at which point the model degenerates into a traditional GCN model. Taking GCN as an example, its space complexity is $O(LND)$, time complexity is $O(L||A||_0D + LND^2)$, where L is the number of model layers, N is the number of nodes, and D is the number of hidden channels. From time and space complexity, it can be seen that the time and space costs of GCN are proportional to the number of model layers. MSFGCN retains only a few convolutional layers when testing, but retains all training parameters during the training phase, so the performance of the model during testing can approach deep graph convolution models, but the time and space costs are close to shallow graph convolution models.

4 Pragmatic competence test based on MSFGCN and multimodal feature fusion

4.1 Overview of the pragmatic competence test model

This paper proposes a pragmatic ability test method based on MSFGCN and multimodal feature fusion. This method mainly consists of two modules: a multimodal fusion module and a long-range sentiment fusion module. The multimodal fusion module consists of three bidirectional fusion modules (BFM) for fusing unimodal features. The long-range sentiment fusion module is used to build sentence information, capture long-range speaker information, and fuse these two types of information. During the sentence information construction process, the edges in the graph are enhanced through a knowledge base.

Figure 1 Pragmatic competence assessment model based on MSFGCN and multimodal feature fusion (see online version for colours)



In the pragmatic ability test task, a set of pragmatic information consists of N consecutive utterances, defining the dialogue as sequence $U = \{u_1, u_2, \ldots, u_N\}$ where u_i represents the i^{th} utterance in the dialogue. A set of dialogues corresponds to M speakers, defining the speakers as sequence $S = \{s_1, s_2, \ldots, S_M\}$, where s_i represents the speaker corresponding to statement u_i. Utterance u_i contains text t, audio a, and visual v modalities, whose feature representations are denoted as u_i^t, u_i^a, and u_i^v. In this paper, $U_t = \left\{u_1^t, u_2^t, \ldots, u_N^t\right\}$, $U_t = \left\{u_1^a, u_2^a, \ldots, u_N^a\right\}$, $U_v = \left\{u_1^v, u_2^v, \ldots, u_N^v\right\}$ represent the text, audio, and visual modality sequences of all statements in the entire dialogue. The objective of the pragmatic ability test task is to predict the ability label of each utterance u_i based on predefined ability categories from these modality features. To this end, the model needs to comprehensively utilise the contextual information of pragmatic information, the multimodal features of utterances, and the information of the speakers to achieve accurate classification of the pragmatic ability of each speaker.

4.2 Bidirectional integration module

To avoid the loss and confusion of information that may occur during one-time fusion, the bidirectional fusion module gradually introduces multimodal information into the model, helping the module to fully utilise the information of different modalities at each stage, performing more refined feature extraction and interaction to improve the fusion effect and overall performance of the model. For the input three modality features m_1, m_2, and m_3, where $m_1 = U_t$, $m_2 = U_a$, $m_3 = U_v$. This module is responsible for first forwardly fusing two of the modalities, and then gradually introducing the third modality and performing backward fusion, as shown in Figure 2, where each BFM performs fusion with the third modality as the main input, so during the fusion process, the first two modalities undergo one Transformer encoding (Foumani et al., 2024), while the third modality undergoes two encodings, ultimately forming a comprehensive multimodal feature representation. This module includes two submodules: the forward fusion submodule and the backward fusion submodule.

The forward fusion submodule first takes modalities m_1 and m_2 as inputs and obtains the output o_1 after going through the Transformer encoder, as shown in equation (7).

$$o_1 = Transformer\left(m_1, m_2 \right) \tag{7}$$

Then, o_1 and the third modality m_3 are taken as inputs, and pass through two transformer encoders consecutively to obtain the output o_3, as shown in equation (8).

$$o_3 = Transformer\left(Transformer\left(o_1, m_3\right), m_3\right) \tag{8}$$

The backward fusion submodule first takes m_1 and m_3 as inputs and obtains the output o_4 after passing through two transformer encoders, as shown in equation (9).

$$o_4 = Transformer\left(Transformer\left(m_1, m_3\right), m_3\right) \tag{9}$$

Finally, o_4 and modality m_2 are taken as inputs and pass through the transformer encoder to obtain the output o_5. By fusing o_3 and o_5, the final output representation o_{final}^1 of the bidirectional fusion module is formed, as shown in equation (10), where W_0 are trainable parameters and b is the bias term.

$$o_{final}^{1} = W_0 \left[o_5, o_3 \right] + b \tag{10}$$

Since this paper adopts three BFM to fuse the initial features, and the characteristics of different modality combinations may have different contributions to sentiment classification, a weight adaptive module is designed to dynamically measure the impact of each modality on the final output. This module performs weighted fusion on the multimodal features and finally generates a comprehensive feature representation. Given the output vectors o_{final}^{1}, o_{final}^{2}, and o_{final}^{3} of the three modules, they are processed through an unbiased linear transformation for each output vector's representation, and the softmax operation is applied to the processed multimodal representation to obtain the weight $n_{softmax}$ of each output, as shown in equation (11), where W_1, W_2 and W_3 represent weights corresponding to different modalities, used to adjust the importance of different modalities, is the vector multiplication operation.

$$n_{\text{softmax}} = \text{softmax}\left(\left[W_1 \otimes o_{final}^{1}, W_2 \otimes o_{final}^{2}, W_3 \otimes o_{final}^{3} \right] \right) \tag{11}$$

Use the weights $n_{softmax}$ to weight and sum the output vectors to obtain the final fused representation N', as shown in equation (12), where $n^{(i)} = o_{final}^{(i)}$, $\odot$ are element-wise products.

$$N' = \sum_{i=1}^{3} n_{\text{softmax}} n^{(i)} \tag{12}$$

4.3 Long-distance integration module

In the pragmatics test task, accurately capturing and understanding the global context and the speaker's specific information is essential for identifying pragmatic competence. Therefore, the distant sentiment fusion module first utilises a GNN to represent pragmatics information in the graph form to construct sentence information, and it uses MSFGCN to capture the distant context information of the speaker. Finally, the above-mentioned sentence information and speaker information are fused. This paper constructs the pragmatic information of the speaker into an undirected graph $G = (V, E)$, where V represents the nodes of the three modalities in each utterance, and E represents the edges between each pair of relational nodes.

MSFGCN updates the hidden state of the nodes by aggregating the representations of its neighbouring nodes according to the type of connected edges, and introduces a distance-aware attention mechanism to enhance the capability of MSFGCN. The process of updating pragmatic information nodes is shown in equation (13).

$$h_i' = \sigma\left(\sum_{r \in R} c_{i,j} \sum_{j \in N_i^r} x_{ij,r} + W_4 h_i \right) \tag{13}$$

where σ represents the activation function, R represents the set of relations, h_i is the input representation of node v_i, $h_i^{'}$ is the output representation of node v_i, N_i^r belongs to the neighbour set v_i under the relation $r \in R$, W_4 is a trainable parameter, and $c_{i,j}$ is a question-specific normalisation constant, usually assigned as the number of neighbours under the relation *r*.

The node representation $h_i^{(l)}$ of each pragmatic information node v_i is updated through equation (14).

$$h_i^{(l+1)} = (1-\beta_i)\left(\sum_{j\in N(i)} \alpha_{i,j} m_j\right)\beta_i W_6 h_i^{(l)} \tag{14}$$

where $N(i)$ is the set of source nodes connected to the target node i, m_j is the information passed from the source nodes, $\alpha_{i,j}$ is the attention score, β_i is the gating parameter of the residual connection, and W_6 is the mapping weight. Based on the above node update rules, the final representations of all nodes in the pragmatic information can be obtained. The final output representation of the pragmatic information is H.

To learn contextual representations at the speaker level, the MSFGCN model employs a Transformer network to capture the self-dependencies between adjacent utterances of a speaker. Given the fused features n_i of each utterance, the speaker-level contextual representation p_i is computed as shown in equation (15), where $h_{\lambda,j}$ is the j^{th} hidden state of speaker p_λ, derived from the speaker-level transformer network. Finally, the attention mechanism fuses H and p_i to obtain the final fused features H_i.

$$p_i = \text{Transformer}(n_i, h_{\lambda,j}) \tag{15}$$

4.4 Model training and prediction results output

The linear layer processes the fused features H_i extracted by the long-range feature fusion module, then applies the ReLU activation function and a Softmax layer to H_i in order to predict the pragmatic competence corresponding to the utterances.

$$H_i' = \text{ReLU}(W_7 H_i + b_1) \tag{16}$$

$$D_i = \text{Softmax}(W_7 H_i' + b_2) \tag{17}$$

$$\hat{y}_i = \arg\max(D_i) \tag{18}$$

where H_i' represents the features after linear transformation and ReLU activation, W_7 represents the weight matrix during linear transformation, b_1 and b_2 denote the bias vectors during linear transformation, D_i represents the output of the Softmax layer, and $\hat{y}_i$ represents the final predicted pragmatic competence category.

This article uses the cross-entropy function (Ho and Wookey, 2019) to compute the loss $L(\theta)$ as shown in equation (19), where N is the number of pragmatic information samples, $c(i)$ is the number of utterances in the i^{th} dialogue, $p_{i,j}$ is the emotion label probability of the j^{th} utterance in the i^{th} dialogue, is the label of the j^{th} utterance in the i^{th} dialogue, and θ is a trainable parameter.

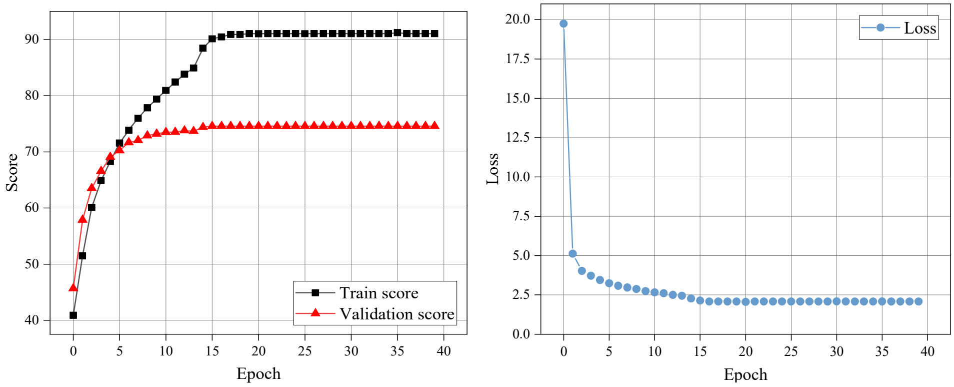
$$L(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{c(i)} \log P_{i,j}\left[y_{i,j}\right] \tag{19}$$

5 Experimental results and performance analysis

5.1 Convergence analysis of the pragmatic competence test model

This article conducts experiments on the public dataset MELD (Rasgado-Toledo et al., 2021). This dataset contains pragmatic data from three modalities: text, audio, and image, covering a variety of competency categories, and is widely used to evaluate the performance of pragmatic competence prediction methods. MELD includes 136,743 statements, with pragmatic competence test results categorised as excellent, good, and weak. The MELD dataset is divided into training, validation, and test sets at a ratio of 8:1:1. The experiment uses the Adam optimiser. To prevent overfitting, dropout is applied after the fully connected layer with a dropout value of 0.1, a learning rate of 0.001, and 100 epochs. All experiments were conducted on a Windows 10 system equipped with an Intel® Core™ i5-6300HQ CPU with 16GB RAM and an NVIDIA GeForce GTX 950M GPU with 4GB VRAM. PyTorch 1.7.0 and CUDA 11.0 toolkits were employed.

Figure 2 Convergence analysis of the MSFGCN training process, (a) training set and validation set scores, (b) training loss analysis (see online version for colours)



(a)

(b)

Figure 2 presents the convergence analysis results of the proposed MSFGCN model during training, divided into the following two parts. As shown in Figure 2(a), during the early training phase, the training set score increases rapidly, rising from approximately 40 points to around 70 points by epoch 5. The growth trend then gradually slows down, stabilising at around 90 points by epoch 20. The validation set score, however, increases relatively steadily. Starting from slightly above 40 points, it gradually rises to approximately 75 points and remains stable. The training set score is higher than the validation set score, with both showing steady improvement. This indicates that the model continuously learns during training and possesses a certain degree of generalisation capability for new data. As shown in Figure 2(b), the training loss analysis depicts the variation of training loss over training epochs. At the start of training, the loss value is relatively high, around 20. It then rapidly decreases, dropping to approximately 5 around epoch 5. Subsequently, the rate of decrease gradually slows down. After epoch 15, the loss value stabilises, settling at around 2.5. This indicates that as training progresses and epochs increase, the model's loss value continuously decreases, ultimately reaching a relatively stable state. This reflects the model's gradual convergence during training.

5.2 Practical competence test accuracy analysis

To further verify the practical competence test accuracy of different models, this paper selects the indicators accuracy, F1, mean absolute error (MAE), root mean square error (RMSE), and ROC curve to evaluate the ECTRANS, TASC, MMDRBN, HFGNN, GATGRU, and MSFGCN models, as shown in Table 1. The accuracy and F1 of MSFGCN are 94.85% and 93.56%, respectively, which are improved by at least 3.46% and 2.5% compared to the baseline models. The MAE and RMSE of MSFGCN are 0.0826 and 0.1153, respectively, which are reduced by at least 21.48% and 22.67% compared to ECTRANS, TASC, MMDRBN, HFGNN, and GATGRU. The MSFGCN model not only optimises the GNN based on a multi-stage adaptive fusion method but also applies the optimised GNN to practical competence tests, significantly improving the test accuracy.

Table 1 Accuracy comparison of different methods in pragmatic ability testing

Model	*Accuracy* (%)	*F1* (%)	*MAE*	*RMSE*
ECTRANS	82.71	81.49	0.1864	0.2037
TASC	85.21	86.34	0.1739	0.1892
MMDRBN	88.63	87.29	0.1423	0.1683
HFGNN	89.21	90.82	0.1152	0.1475
GATGRU	91.39	91.06	0.1052	0.1391
MSFGCN	94.85	93.56	0.0826	0.1153

The ROC curves of different models are shown in Figure 3, with AUC values for ECTRANS, TASC, MMDRBN, HFGNN, GATGRU, and MSFGCN being 0.7852, 0.8236, 0.8401, 0.9029, 0.9374, and 0.9756, respectively. MSFGCN improves by 24.25%, 18.46%, 16.13%, 8.05%, and 4.08% compared to ECTRANS, TASC, MMDRBN, HFGNN, and GATGRU, respectively. ECTRANS uses a transformer model to implement a pragmatic ability test, but the performance of transformer models highly depends on the quality and diversity of the training data. If the training data does not adequately cover different language backgrounds, cultural customs, or social scenes, the model may not accurately evaluate the pragmatic abilities of specific groups. TASC performs pragmatic ability prediction through fine-grained temporal alignment and cross-modal semantic interaction. However, fine-grained temporal alignment requires precise matching of corresponding relationships across modalities on the timeline, but existing datasets often lack such high-precision annotations. MMDRBN involves multimodal data such as speech, text, and body language, and the synchronisation of these data on the timeline is crucial. However, the sampling frequency and start time of different modalities may vary, leading to difficulties in temporal alignment. HFGNN's pragmatic ability prediction model based on heterogeneous graphs integrates multiple types of nodes and complex relational edges, but in pragmatic ability data, certain pragmatic behaviours may have very few samples, leading to sparsity in the corresponding nodes and edges in the heterogeneous graph. The GATGRU pragmatic ability prediction model based on GAT captures complex interaction relationships by dynamically assigning attention weights between nodes. However, the context in pragmatic ability tests may change rapidly, but GAT is typically based on static graph structures, making it difficult to update node features or attention weights in real-time to reflect dynamic changes. MSFGCN improves prediction accuracy by introducing a multimodal fusion module and a long-range feature fusion module.

Figure 3 ROC curve for the pragmatic competence test method (see online version for colours)

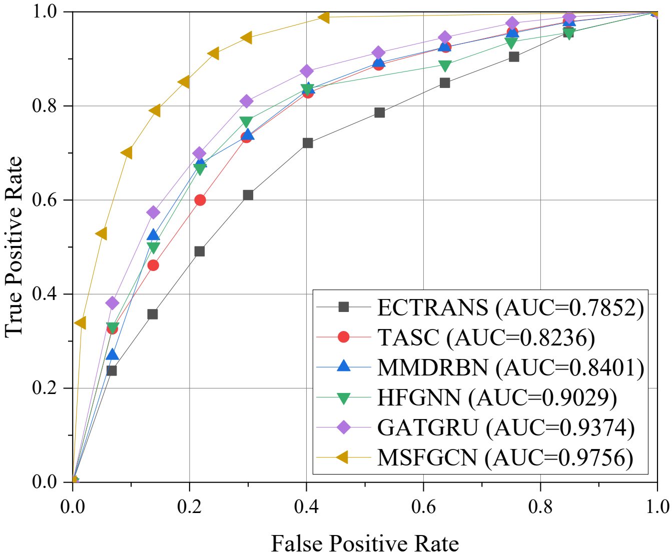
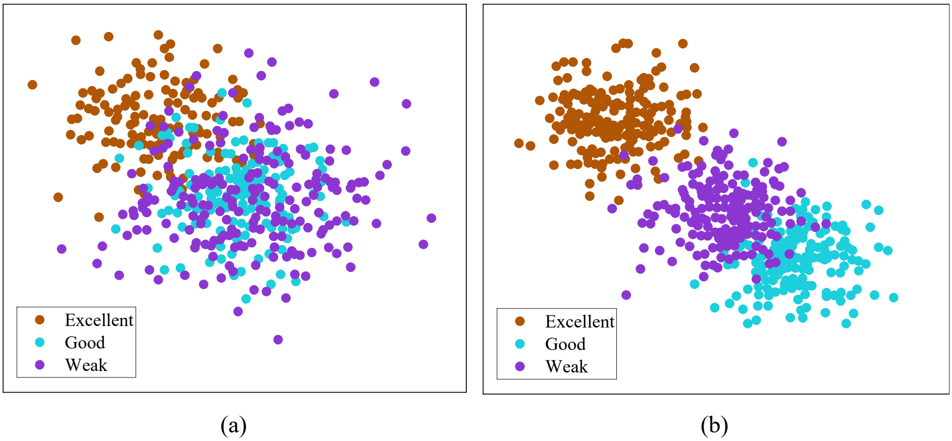


Figure 4 Visualisation results for the three-category task in the pragmatic competence test, (a) distribution of raw pragmatic competence data, (b) distribution of raw pragmatic competence data processed by MSFGCN (see online version for colours)



(a) (b)

Furthermore, Figure 4 shows the visualisation results of the three-class task in the pragmatic ability test task. The three colours represent different types of pragmatic ability categories. Figure 4(a) displays the original visualisation of the data, where feature points of different categories are intertwined and difficult to distinguish clearly. However, in Figure 4(b), the feature points after processing by MSFGCN show a more distinct clustering effect, and clear boundaries are formed between feature points of different pragmatic ability categories. This fully demonstrates that the introduction of the multimodal fusion module and the long-range feature fusion module not only enables the model to fully utilise multimodal information but also significantly improves the model's expressive capacity and classification efficiency in complex pragmatic ability test tasks.

6 Conclusions

Current methods of pragmatic ability tests focus more on the capture of local contexts and often overlook the integration of long-range pragmatic features when handling extended pragmatic information. To address these issues, this paper first proposes a GNN based on multi-stage adaptive fusion, called MSFGCN. MSFGCNN decomposes the deep GNN model into a multi-stage training framework, where each stage contains several feature extraction layers. The main function of the deep learning module based on multi-stage training is to gradually incorporate deep graph information into a shallow model for training a more powerful shallow model. Subsequently, a pragmatic ability prediction model based on MSFGCN and multimodal feature fusion is proposed. The model consists of: a multimodal fusion module and a long-range sentiment fusion module. The multimodal fusion module consists of three BFM units for fusing unimodal features. The long-range feature fusion module is used to construct sentence information, capture long-range speaker pragmatic information, and fuse these two types of information. During the sentence information construction process, the edges in the graph are enhanced through a knowledge base. Experimental results show that the pragmatic ability test accuracy and AUC of the proposed model are 94.85% and 0.9756, respectively, which are improved by at least 3.46% and 4.08% compared to baseline models, thereby positively promoting the development of language teaching and natural language processing applications.

Although the proposed model has achieved high accuracy in pragmatic ability tests, the current method of constructing pragmatic data into graph structures still has room for improvement in terms of node and edge definitions. Future research can explore more refined node partitioning strategies, for example, not only using words and sentences as nodes, but also incorporating abstract concepts such as pragmatic rules and pragmatic strategies as nodes, so that the graph structure can more comprehensively and accurately reflect the complexity of pragmatic relationships. At the same time, the method for calculating edge weights can also be further optimised, with a dynamic weight mechanism introduced to adjust edge weights in real-time according to different pragmatic scenarios and tasks, thereby better capturing the dynamic associations between pragmatic information.

Acknowledgements

This work is supported by the National Social Science Fund of China 'The Relationship Between Pragmatic Ability and Executive Function in Preschool Children and Its Applications' (No. 23KY1026).

Declarations

All authors declare that they have no conflicts of interest.

References

Ai, W., Shou, Y., Meng, T. and Li, K. (2024) 'DER-GCN: dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 36, No. 3, pp.4908–4921.

Alsuhaibani, Z. (2022) 'Developing EFL students' pragmatic competence: the case of compliment responses', *Language Teaching Research*, Vol. 26, No. 5, pp.847–866.

Chen, J. (2023) 'A novel model for language training assessment based on data mining and Bayesian network', *Tehnički Vjesnik*, Vol. 30, No. 3, pp.771–778.

Chowanda, A., Sutoyo, R. and Tanachutiwat, S. (2021) 'Exploring text-based emotions recognition machine learning techniques on social media conversation', *Procedia Computer Science*, Vol. 179, pp.821–828.

Dai, H. and Zhao, T. (2022) 'Intelligent analysis strategy of pragmatic failure in cross-cultural communication based on convolution neural network', *Mobile Information Systems*, Vol. 20, No. 3, pp.78–91.

Eragamreddy, N. (2025) 'The impact of AI on pragmatic competence', *Journal of Teaching English for Specific and Academic Purposes*, Vol. 8, pp.169–189.

Fathi, M.J., Kafipour, R., Kashefian-Naeeini, S. and Shahsavar, Z. (2025) 'The impact of reflective teaching on EFL learners through implicit and explicit pragmatic competence instructions', *Reflective Practice*, Vol. 26, No. 2, pp.292–310.

Flores, M.J., Gámez, J.A. and Martínez, A.M. (2014) 'Domains of competence of the semi-naive Bayesian network classifiers', *Information Sciences*, Vol. 260, pp.120–148.

Foumani, N.M., Tan, C.W., Webb, G.I. and Salehi, M. (2024) 'Improving position encoding of transformers for multivariate time series classification', *Data Mining and Knowledge Discovery*, Vol. 38, No. 1, pp.22–48.

He, Z., Li, W. and Yan, Y. (2022) 'Modeling knowledge proficiency using multi-hierarchical capsule graph neural network', *Applied Intelligence*, Vol. 52, No. 7, pp.7230–7247.

Ho, Y. and Wookey, S. (2019) 'The real-world-weight cross-entropy loss function: modeling the costs of mislabeling', *IEEE Access*, Vol. 8, pp.4806–4813.

Kentmen, H., Debreli, E. and Yavuz, M.A. (2023) 'Assessing tertiary Turkish EFL learners' pragmatic competence regarding speech acts and conversational implicatures', *Sustainability*, Vol. 15, No. 4, pp.38–52.

Kim, B., Chung, K., Lee, J., Seo, J. and Koo, M-W. (2019) 'A Bi-LSTM memory network for end-to-end goal-oriented dialog learning', *Computer Speech and Langu*age, Vol. 53, pp.217–230.

Li, W., Chen, Q., Gu, G. and Sui, X. (2025) 'Object matching of visible–infrared image based on attention mechanism and feature fusion', *Pattern Recognition*, Vol. 158, pp.11–20.

Li, Z., Lin, W. and Zhang, Y. (2023) 'Drive-by bridge damage detection using Mel-frequency cepstral coefficients and support vector machine', *Structural Health Monitoring*, Vol. 22, No. 5, pp.3302–3319.

Ma, G., Yang, X., Zhang, B. and Shi, Z. (2016) 'Multi-feature fusion deep networks', *Neurocomputing*, Vol. 218, pp.164–171.

Parola, A., Gabbatore, I., Berardinelli, L., Salvini, R. and Bosco, F.M. (2021) 'Multimodal assessment of communicative-pragmatic features in schizophrenia: a machine learning approach', *Nature Partner Journal Schizophrenia*, Vol. 7, No. 1, pp.28–43.

Planques, V. and Julián, M. (2018) 'English language learners' spoken interaction: what a multimodal perspective reveals about pragmatic competence', *System*, Vol. 77, pp.80–90.

Prasatyo, B.A., Ali, H.V. and Hidayati, D. (2023) 'Current studies on pragmatics competence in EFL learning context: a review', *Jurnal Sinestesia*, Vol. 13, No. 2, pp.985–994.

Rasgado-Toledo, J., Lizcano-Cortés, F., Olalde-Mathieu, V.E., Licea-Haquet, G., Zamora-Ursulo, M.A., Giordano, M. and Reyes-Aguilar, A. (2021) 'A dataset to study pragmatic language and its underlying cognitive processes', *Frontiers in Human Neuroscience*, Vol. 15, pp.66–80.

Salamanti, E., Park, D., Ali, N. and Brown, S. (2023) 'The efficacy of collaborative and multimodal learning strategies in enhancing English language proficiency among ESL/EFL Learners: a quantitative analysis', *Research Studies in English Language Teaching and Learning*, Vol. 1, No. 2, pp.78–89.

Sinclair, J., Jang, E.E. and Rudzicz, F. (2021) 'Using machine learning to predict children's reading comprehension from linguistic features extracted from speech and writing', *Journal of Educational Psychology*, Vol. 113, No. 6, pp.24–35.

Verma, A.K., Saxena, R., Jadeja, M., Bhateja, V. and Lin, J.C-W. (2023) 'Bet-GAT: an efficient centrality-based graph attention model for semi-supervised node classification', *Applied Sciences*, Vol. 13, No. 2, pp.84–107.

Wahlster, W. (2023) 'Understanding computational dialogue understanding', *Philosophical Transactions of the Royal Society A*, Vol. 381, No. 22, pp.49–63.

Yan, B-C. and Chen, B. (2024) 'An effective hierarchical graph attention network modeling approach for pronunciation assessment', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 32, pp.3974–3985.

Zainal, A.G., Misba, M., Pathak, P., Patra, I., Gopi, A., El-Ebiary, Y.A.B. and Prema, S. (2024) 'Cross-cultural language proficiency scaling using transformer and attention mechanism hybrid model', *International Journal of Advanced Computer Science and Applications*, Vol. 15, No. 6, pp.74–89.

Zhou, Y., Zheng, H., Huang, X., Hao, S., Li, D. and Zhao, J. (2022) 'Graph neural networks: Taxonomy, advances, and trends', *ACM Transactions on Intelligent Systems and Technology* (*TIST*), Vol. 13, No. 1, pp.1–54.