Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 17 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i17.57857

PAPER

Emotion-Aware Mental Health Intervention Strategies on Mobile Computing Platforms

Yang Li[1], Lihua Peng[2] (✉)

[1]Centre for Mental Health Education, China People's Police University, Langfang, China

[2]Periodical Department, Langfang Normal University, Langfang, China

penglihuamail@126.com

ABSTRACT

With the rapid advancement of mobile computing technologies and the continuous growth of global mobile device users, smartphones and other mobile terminals have created new possibilities for the widespread delivery of mental health services. At the same time, the increasing psychological pressure in modern society—especially in the post-pandemic era—has led to a surging demand for mental health interventions. Leveraging mobile computing platforms for accurate emotion recognition and effective intervention has thus become a critical area of research. However, current studies show that unimodal emotion recognition models achieve less than 60% accuracy in complex scenarios, while multimodal fusion methods often overlook the impact of spatiotemporal context, resulting in significant accuracy degradation in cross-regional emotion recognition. Furthermore, existing approaches lack effective strategies to address data sparsity and noise in mobile environments. This study focuses on emotion analysis in mental health service dialogues conducted via mobile networks. We propose a multimodal fusion module and a long-distance emotion fusion module. The former integrates multi-source data—including text, voice, and facial expressions—for comprehensive emotion capture, while the latter constructs a cross-regional emotional feature mapping mechanism to incorporate spatiotemporal contextual information. The contribution of this study lies in overcoming the limitations of existing models in terms of recognition accuracy and adaptability to diverse scenarios, thereby offering a feasible technical framework for mental health interventions on mobile computing platforms and advancing the intelligent development of digital mental health services.

KEYWORDS

mobile computing platform, emotion recognition, mental health intervention strategies, multimodal fusion module, long-distance emotion fusion module

Li, Y., Peng, L. (2025). Emotion-Aware Mental Health Intervention Strategies on Mobile Computing Platforms. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(17), pp. 101–114. https://doi.org/10.3991/ijim.v19i17.57857

Article submitted 2025-05-24. Revision uploaded 2025-07-24. Final acceptance 2025-07-29.

1 INTRODUCTION

With the rapid development of mobile computing technology, mobile terminals such as smartphones and wearable devices [1–3] have become an indispensable part of people's daily lives. According to statistics, the number of global mobile device users continues to rise. Relying on its portability, real-time capability, and popularity [4–6], the mobile platform provides new possibilities for the wide coverage of mental health services. At the same time, the psychological pressure faced by people in modern society is increasing day by day [7, 8], and mental health problems are becoming increasingly prominent. In the post-pandemic era, the public's demand for mental health intervention has risen sharply [9, 10]. Against this background, how to utilize the technical advantages of mobile computing platforms to achieve accurate emotion recognition and carry out effective mental health intervention accordingly has become an important issue of common concern in academia and industry.

Emotion, as an important external manifestation of psychological state [11], its accurate recognition is a key prerequisite for carrying out mental health intervention [12]. The study on emotion recognition and mental health intervention strategies on mobile computing platforms has important theoretical and practical significance. However, the existing research on emotion recognition and mental health intervention based on mobile computing platforms still has obvious defects and shortcomings in methodology. Many studies rely only on unimodal data for emotion recognition [12–15], which makes it difficult to comprehensively capture users' complex emotional states. For example, the study in literature [16] shows that the accuracy of unimodal emotion recognition in complex scenarios is generally less than 60%, which is difficult to meet the actual intervention needs. In addition, although some studies adopt multimodal fusion methods [17, 18], they often ignore the influence of spatiotemporal context on emotional expression when processing long-distance emotion information. Literature [19] points out that existing models show a significant decrease in recognition accuracy when dealing with cross-regional and cross-cultural emotional data, which greatly weakens the effect of remote mental health intervention based on mobile platforms. At the same time, most existing studies lack effective strategies to deal with data sparsity and noise in mobile environments, which further limits the accuracy and adaptability of intervention strategies.

Based on the above background and problems, this paper focuses on "emotion analysis of mental health service dialogues based on mobile networks," aiming to construct a more accurate emotion recognition model to support effective mental health intervention. The main research content of the paper includes the design and implementation of two core modules: a multimodal fusion module and a long-distance emotion fusion module. Among them, the multimodal fusion module integrates multi-source data such as text, voice, and facial expressions and realizes comprehensive capture of users' emotions through deep learning methods; the long-distance emotion fusion module focuses on considering spatiotemporal contextual information and improves the model's ability to process long-distance emotional information by constructing a cross-regional emotion feature mapping mechanism. The value of this study lies in the innovative combination of multimodal fusion and long-distance emotion processing, which is expected to break through the limitations of existing research in emotion recognition accuracy and scenario adaptability, provide more feasible technical

solutions for mental health intervention on mobile computing platforms, promote the intelligent development of digital mental health services, and provide strong support for the realization of personalized and precise mental health intervention.

2 EMOTION ANALYSIS OF MENTAL HEALTH SERVICE DIALOGUES BASED ON MOBILE NETWORKS

This paper focuses on the construction of an emotion analysis model for mental health service dialogues, based on multimodal dialogue data under mobile network environments, and adopts a "hierarchical fusion–context enhancement" design idea: First, through three bidirectional fusion submodules in the multimodal fusion module, the unimodal features such as text semantics, voice prosody, and facial expressions are fused across dimensions. With the help of a bidirectional transmission mechanism, feature complementation and reinforcement between different modalities are realized, solving the problem of one-sided information in a single modality. At the same time, the long-distance emotion fusion module constructs a dialogue semantic graph based on graph neural networks, treating each round of dialogue as a node. By integrating a psychological knowledge base and dialogue behavior rules, the semantic association between nodes is enhanced, thereby capturing the temporal transfer patterns and cross-round influences of speaker emotions in multi-turn dialogues. Then, through graph convolution operations, long-distance speaker emotion information is aggregated and mapped.

2.1 Multimodal fusion module

Aiming at the multimodal emotion analysis needs of mental health service dialogues, this paper adopts a "progressive modality interaction–dominant feature reinforcement" idea to design the bidirectional fusion module. For the three core modality features in mental health dialogues—text semantics l_1, voice prosody l_2, and facial micro-expressions l_3—the module first uses a forward fusion submodule to perform cross-modal feature interaction on any two modalities after Transformer encoding. The attention mechanism is used to capture the association between semantic emotion and voice emotion. Then, taking the third modality as the dominant one, a reverse fusion submodule is introduced. Through two rounds of Transformer encoding, the weight of facial expression features in cross-modal fusion is enhanced, enabling non-verbal emotional cues like facial micro-expressions to map bidirectionally with the other two modality features at a deeper level (see Figure 1). This phased fusion mechanism is particularly suited to the multi-level emotional transmission characteristics of "verbal expression–paralinguistic cues–nonverbal signals" in mental health dialogues. It avoids emotional feature confusion caused by one-time fusion and ensures that key emotional modalities can be preferentially captured in multi-turn dialogues between counselor and user through the double encoding mechanism of the dominant modality. Specifically, let the forward fusion submodule take modalities l_1 and l_2 as inputs, and obtain the output $p_1 \in E^{V \times f}$ after the Transformer encoder:

$$p_1 = TR(l_1, l_2) \tag{1}$$

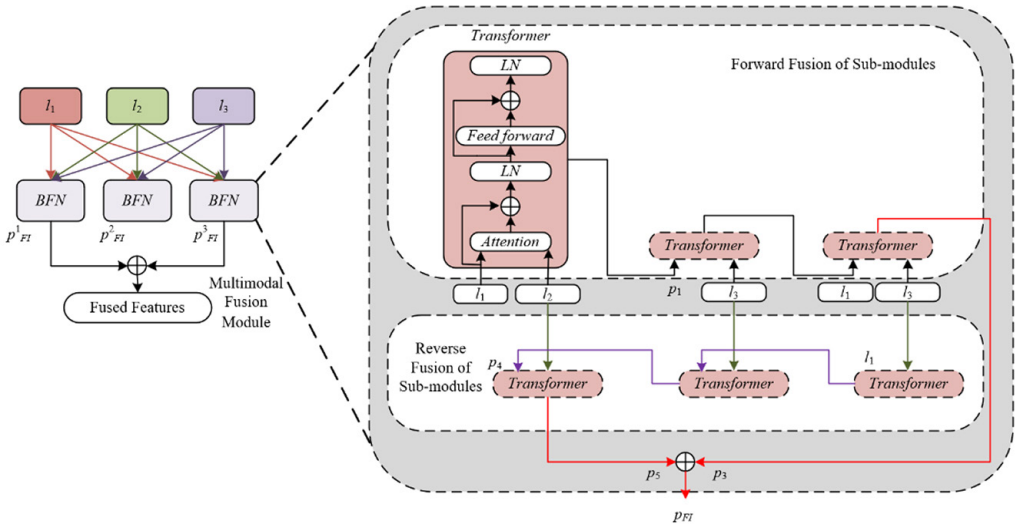


Fig. 1. Multimodal fusion module architecture

Then, take p_1 and the third modality l_3 as input, and obtain the output p_3 after two Transformer encoders:

$$p_3 = TR(TR(p_1, l_3), l_3) \tag{2}$$

The reverse fusion submodule takes l_1 and l_3 as input and obtains the output p_4 after two Transformer encoders:

$$p_4 = TR(TR(p_1, l_3), l_3) \tag{3}$$

Finally, take p_4 and modality l_2 as input, and obtain the output p_5 through a Transformer encoder. Suppose the trainable parameters are denoted as Q_0 and the bias term as y; then fuse p_3 and p_5 as the final output p_{FI}^1 of the bidirectional fusion module:

$$p_{FI}^1 = Q_0[p_5, p_3] + y \tag{4}$$

To adapt to the hierarchical emotional transmission pattern of "verbal expression–paralinguistic signals–nonverbal cues" in mental health dialogues, this paper uses three bidirectional fusion modules to respectively process modality combinations of text–voice, voice–expression, and text–expression. Each module realizes cross-modal feature interaction based on a "forward–reverse" phased fusion mechanism, i.e., bidirectional mapping of text semantics and voice prosody, and feature reinforcement between voice emotion and facial micro-expression. The weight adaptive module dynamically calculates the contribution weight of each modality combination through bias-free linear transformation and *softmax* activation based on the output vectors p_{FI}^1, p_{FI}^2, and p_{FI}^3 of the three modules. Through the above design, when the user describes traumatic experiences, the weights of voice trembling and pupil contraction will be dynamically increased by the adaptive module, while the weight of text semantics is relatively reduced, thus avoiding the problem where key emotional signals are masked by verbal expression in one-time fusion. The final fused representation generated by weighted summation not only retains the complementary information of each modality combination but also dynamically adjusts the feature weights in real time according to the modality specificity of emotional outbursts in mental health dialogue scenarios. Specifically, let the weights corresponding to

different modalities be denoted as Q_1, Q_2, Q_3. The weight matrix Q_1, Q_2, Q_3 and the corresponding output vectors p_{FI}^1, p_{FI}^2, and p_{FI}^3 are used to perform vector product operation. Then, the weight v_{SM} of each output is given by the following formula:

$$v_{SM} = \text{softmax}([Q_1 \otimes p_{FI}^1, Q_2 \otimes p_{FI}^2, Q_3 \otimes p_{FI}^3]) \tag{5}$$

Assume the element-wise product is denoted as $v^{(u)} = p_{FI}^{(u)}$, and the final fused representation V' is obtained by weighting and summing the output vectors.

$$V' = \sum_{u=1}^{3} v_{SM} \otimes v^{(u)} \tag{6}$$

2.2 Long-distance emotion fusion module

Considering the "distributed interaction" characteristics of mental health dialogues under mobile networks and targeting the long-distance emotional dependency characteristics of mental health service dialogues in mobile network environments, this paper designs a long-distance emotion fusion module. A dialogue semantic graph is constructed based on the mobile network communication protocol, treating each round of dialogue as a node. The edge weights between nodes are enhanced using the temporal sequence information transmitted over the mobile network and a psychological knowledge base, thus constructing sentence information that contains contextual semantics. At the same time, a *Transformer* encoder is used to model the historical dialogue sequences of the speaker in the mobile network, capturing the emotional transfer patterns in cross-round utterances through a self-attention mechanism. In practical application scenarios, when the counselor and user engage in multi-turn dialogues over mobile networks, the graph neural network can update the dialogue graph structure in real time to adapt to the dynamic transmission of mobile data, while the *Transformer* can extract the long-term emotional features of the speaker from the fragmented dialogues transmitted through the mobile network. Finally, a gating mechanism is used to fuse sentence-level semantic features and cross-round speaker emotional features, which not only retains the contextual information of single-round dialogues but also strengthens the temporal correlation of long-distance speaker emotions in the mobile network environment.

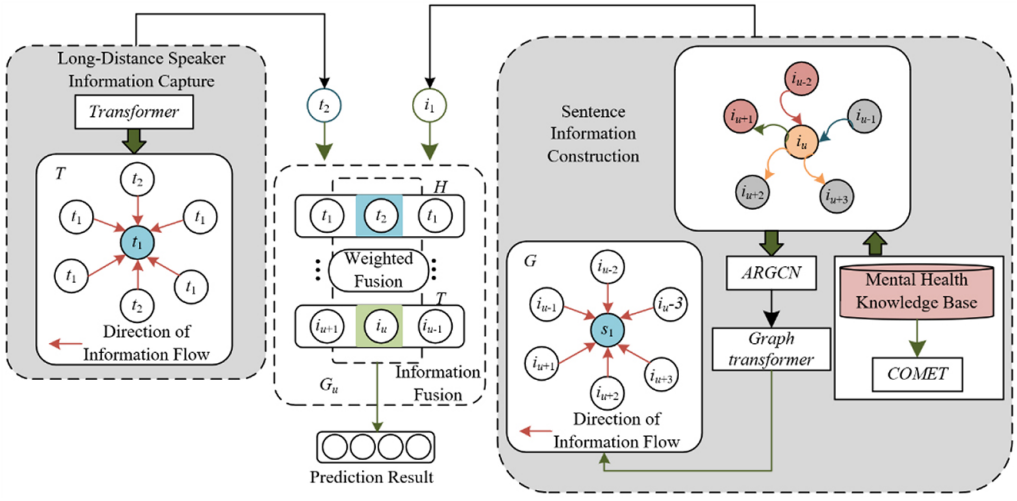


Fig. 2. Long-distance emotion fusion module architecture

Based on the communication characteristics of mobile networks and the interaction logic of mental health dialogues, this paper first constructs a directed dialogue graph representation in the long-distance emotion fusion module as $H = (N, \gamma, E, X)$ (see Figure 2). Each sentence i_u in the dialogue is taken as a graph vertex n_u, and the initial feature g_u of the node is set to the output V'_u from the multimodal fusion module, which inherently contains cross-modal conversation context information such as text, speech, and facial expression. The edge $r_{u,k} = (n_u, e_{uk}, n_k)$ realizes semantic enhancement through the definition of three core relationship types. When the utterances $t_u \rightarrow t_k$ are from the same speaker, the edge type e_{uk} is defined as *aWant* or *aIntent*; when the utterances are from different speakers, it is defined as *pWant*. The COMET model is used to generate semantic representations for the corresponding relationship types to strengthen the edge attribute X, thus embedding professional knowledge such as "psychological intention–action association" into the graph structure. At the same time, based on the temporal characteristics of mobile network transmission, a local context window matching the fragmented dialogue features in mobile communication is constructed by limiting the association scope of the current utterance with the past O and future D utterances through hyperparameters O and D, ultimately forming a directed graph model that can capture dynamic temporal dependencies in mobile networks and represent the professional logic of "intention prediction–action guidance–cross-subject influence" in mental health dialogues.

When constructing sentence information in the long-distance fusion module, this paper considers the dynamic characteristics of mobile network communication and the transmission pattern of psychological states in mental health dialogues. Specifically, an attention-based relational graph convolutional network (ARGCN) is used to update the node states in the directed dialogue graph. Neighbor node features are weighted and aggregated based on edge types. For example, when the edge type is *oWant*, the transmission of features reflecting the impact of actions between different speakers is strengthened. At the same time, a distance-aware attention mechanism is introduced to dynamically adjust the weights of neighbor nodes according to the context window limited by mobile network hyperparameters O and D, solving the temporal dependency problem in fragmented mobile dialogues. Specifically, let the activation function be denoted by δ, the relationship set by E, the input representation of node n_u by $g_u = V'_u$, and the output representation by g'_u. Let V_u^e be the set of neighbors of n_u under relation e, Q_4 be the trainable parameters, and $z_{u,e}$ be the task-specific normalization constant. The node update process for the dialogue graph is given by the following expression:

$$g'_u = \delta\left(\sum_{e \in E} z_{u,k} \sum_{e \in E} a_{uk,e} + Q_4 g_u\right) \tag{7}$$

Then, to adapt to the temporal volatility of dialogue data in mobile networks and consider the edge representations enhanced by psychological professional knowledge to capture the emotional state evolution trajectories in dialogues, this paper applies L layers of GraphTransformer for multi-layer feature propagation. The self-attention mechanism is used to capture long-distance emotional associations across dialogue turns, such as the gradually revealed depressive tendency in users' multi-round dialogues. In each GraphTransformer layer, the node representation is updated through Equation (8), fusing the local interaction information aggregated by ARGCN with the global semantic dependencies across nodes. Thus, the sentence information contains not only the multimodal emotional features of single-turn dialogues but also embeds the unique "intention–action" transmission logic of mental health dialogues. Specifically, let the set of source nodes connected to the target node

u be denoted by $V(u)$, the information transmitted from the source node be denoted by l_k, the attention score by $\beta_{u,k}$, the gating parameter for residual connection by α_u, and the mapping weight by Q_6. The node representation $g_u^{(m)}$ of each utterance node n_u is updated by the following equation:

$$g_u^{(m+1)} = (1-\alpha_u)\left(\sum_{k\in V(u)} \beta_{u,k} l_k\right) + \alpha_u Q_6 g_u^{(m)} \tag{8}$$

Based on the above node update rules, the final representation of all nodes in the dialogue can be obtained. The final output representation of the dialogue is denoted as $H \in R^{du\times du}$, $G \in E^{fi\times fi}$.

Targeting the "asynchronous interaction" characteristics of psychological counseling in mobile network environments and considering the need to extract long-term emotional features of speakers from fragmented dialogue segments, this paper adopts a dual-layer capture method of "sequence modeling–self-dependency reinforcement" to obtain the temporal characteristics of mobile network communication and the dynamic evolution patterns of speaker emotions in mental health dialogues. Specifically, the Transformer network is used to model the utterance sequence of the same speaker, taking the fused utterance features $v_u = V_u'$ output from the multimodal fusion module as input. The self-attention mechanism is used to mine the temporal dependency relationships between adjacent utterances, and position encoding is used to embed the temporal information of mobile network transmission, solving the problem of long-distance dependency breakage caused by fragmented dialogue transmission. When calculating the speaker-level context representation o_u, the Transformer uses the multi-head attention mechanism to capture emotional association patterns at different time spans in parallel. For example, for the common phenomenon of the "emotional latency period" in mental health dialogues, the model will automatically assign higher attention weights to later explosive utterances, enhancing the ability to capture emotional mutation points. Assume that all utterances of speaker o_η are denoted by I_η, and the k-th hidden state of speaker o_η is denoted by $g_{\eta,k}$. The following formula gives the computation of o_u:

$$o_u = TR(v_u, g_{\eta,k}), k \in \left[1, \left|I_\eta\right|\right] \tag{9}$$

Furthermore, considering the temporal dynamics of mental health dialogues in mobile network environments and the cross-round correlation of speaker emotions, this paper chooses to perform information fusion within the long-distance fusion module. Specifically, the current sentence feature A_u is obtained by aggregating the previous v_{PA} utterances, while the speaker feature T_u is formed by integrating all utterance features of the same speaker from the current sentence to the end of the dialogue, such that A_u and T_u respectively carry the local contextual dependency of the dialogue and the long-term emotional trajectory of the speaker. Assuming all utterances in the dialogue are denoted by I_η, then:

$$T_u = ARGCN(T_u, T_k), k \in \left[u, \left|I_\eta\right|\right] \tag{10}$$

Furthermore, the semantic similarity $SIM(A_u, T_u)$ between A_u and T_u is calculated, and the correlation between A_u and all speaker features T_u is further measured, which serves as the attention weight μ_{uk} during Transformer fusion. In mental health dialogue scenarios, when a user's utterance contains semantic

contradiction, the model strengthens the attention weight between this utterance and the historical emotional burst points of the speaker through similarity computation so that Transformer preferentially activates the relevant cross-round emotional features during fusion. Assuming the number of utterances in the dialogue is V, the calculation is as follows:

$$SIM(A_u, T) = \sum_{u=1}^{V} SIM(A_u, T_u) \tag{11}$$

The final fusion result G_u is given by the following equation:

$$G_u = TR(A_u, T, \mu_{uk}) \tag{12}$$

This fusion mechanism is particularly capable of capturing local temporal dependencies in mobile data transmission through A_u and aggregating speaker-level features through T_u, achieving a hierarchical fusion of "instant emotional signals–long-term psychological patterns" via dynamic weights. The final generated feature vector G_u contains both the multimodal emotional information of the current sentence and the speaker's emotional evolution logic throughout the mobile dialogue cycle, providing a structured feature representation that integrates spatiotemporal context and psychological professional knowledge for subsequent mental health intervention strategies.

During the model training phase, this paper chooses to map the feature vector G_u output from the long-distance fusion module to the emotion category space through a linear layer and uses the ReLU activation function to enhance the model's representation capability for nonlinear emotional patterns in mental health dialogues. A Softmax layer then generates a probability distribution conforming to the psychological emotion classification system. Suppose the feature after linear transformation and ReLU activation is represented as G'_u, the weight matrix in the linear transformation is Q_7, the bias vectors are y_1 and y_2, the output of the Softmax layer is F_u, and the finally predicted emotional category is $\hat{b}$, then the computation is as follows:

$$G'_u = \text{ReLU}(Q_7 G_u + y_1) \tag{13}$$

$$F_u = \text{softmax}(Q_7 G'_u + y_2) \tag{14}$$

$$\hat{b}_u = \text{argmax}(F_u) \tag{15}$$

During training, the cross-entropy function is used as the optimization objective. Considering the fragmented nature of mobile network dialogue data, a temporal weighting mechanism is introduced in the loss function computation. Higher weights are assigned to key emotional frames consistent with the logic of mental health intervention, guiding the model to prioritize learning emotional features in the dialogue that have psychological intervention value. Assume the number of all dialogues is V, the number of utterances in the u-th dialogue is $z(u)$, the probability of the emotional label for the k-th utterance in the u-th dialogue is $O_{u,k}$, the label of the k-th utterance in the u-th dialogue is $b_{u,k}$, and the trainable parameters are denoted by ϕ. The loss function is expressed as:

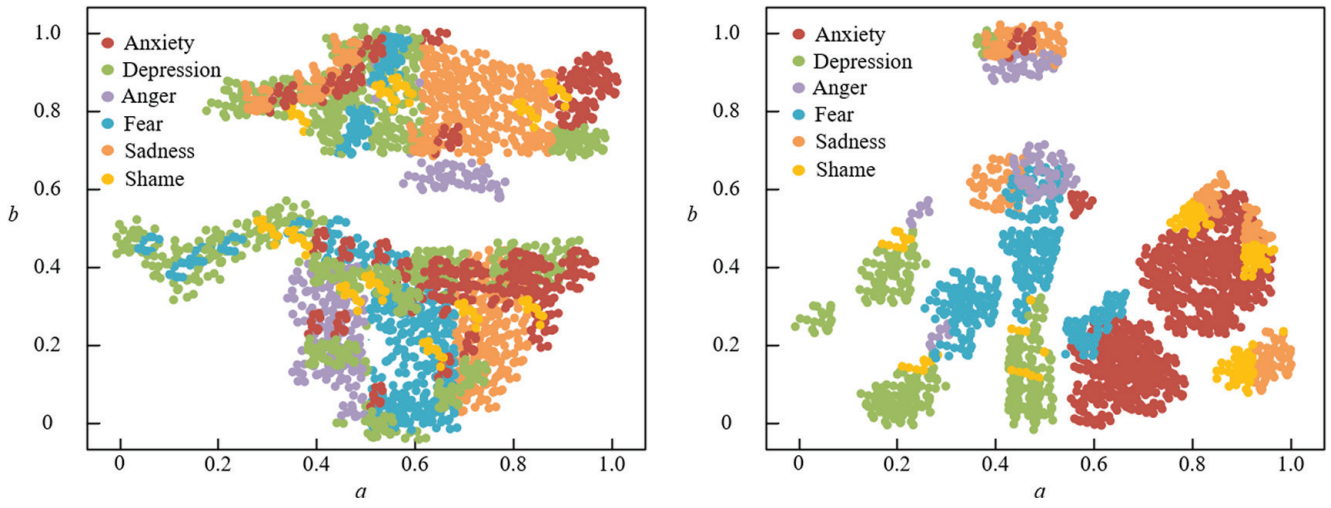
$$LOSS(\varphi) = -\sum_{u=1}^{V} \sum_{k=1}^{z(u)} \log O_{u,k}[b_{u,k}] \tag{16}$$

3 EXPERIMENTAL RESULTS AND ANALYSIS

As shown in Table 1, the ablation experiment results indicate that on both the training and testing sets, the complete model demonstrates better performance on most of the weighted average F1 and Acc indicators. Taking the test set as an example, the complete model outperforms the models with submodules removed on certain indicators. Comparing the cases of removing the multimodal fusion submodule and the long-distance emotion fusion module, the complete model integrates three core modality features: text semantics l_1, voice prosody l_2, and facial micro-expressions l_3, as well as multimodal fusion and long-distance emotion fusion mechanisms, which provides an advantage in the accuracy of emotion recognition. This indicates that the mobile network-based emotion analysis method for mental health service dialogues proposed in this paper effectively improves the model performance in emotion recognition through the multimodal fusion and long-distance emotion fusion modules.

Table 1. Results of ablation experiments

Dataset	Ablated Module	Weighted Average F1				Acc			
		l_1+l_2	l_1+l_3	l_2+l_3	$l_1+l_2+l_3$	l_1+l_2	l_1+l_3	l_2+l_3	$l_1+l_2+l_3$
Training Set	Remove Reverse Fusion Submodule	67.25	62.31	61.25	68.32	67.89	62.58	61.25	68.32
	Remove Forward Fusion Submodule	66.32	64.25	63.45	68.25	66.32	63.21	61.48	68.52
	Remove Long-distance Emotion Fusion Module	66.89	63.58	54.89	66.54	66.52	63.48	54.32	66.31
	Complete Model	68.52	64.25	62.31	72.32	68.52	64.21	62.36	71.56
Test Set	Remove Reverse Fusion Submodule	64.21	64.58	32.58	64.21	65.41	65.26	47.89	65.32
	Remove Forward Fusion Submodule	64.59	64.23	37.56	64.58	65.32	66.23	45.32	66.98
	Remove Long-distance Emotion Fusion Module	63.21	63.51	45.21	63.26	64.58	65.89	52.31	65.24
	Complete Model	64.56	64.56	41.23	64.59	65.32	66.32	48.69	66.23



a) Initial Feature Distribution **b) Feature Distribution Learned by the Model**

Fig. 3. t-SNE visualization of different emotion categories

According to the *t-SNE* visualization results in Figure 3, the feature distribution of different emotion categories in the initial state is relatively chaotic, with feature points of various emotions intertwined and boundaries blurred, indicating that the original multi-source data did not form clearly distinguishable emotional feature representations. After model learning, the feature distribution of various emotions shows an obvious clustering trend, with emotional feature points relatively aggregated and category boundaries clear. This demonstrates that the proposed multimodal fusion module and long-distance emotion fusion module play a key role. The multimodal fusion integrates multi-source data such as text, voice, and facial expressions, and the long-distance emotion fusion constructs cross-regional feature mappings considering the spatiotemporal context. Together, they enable the model to learn more distinguishable and accurate emotional features, strongly validating the effectiveness of the proposed mobile network-based mental health service dialogue emotion analysis method in improving emotional feature representation and recognition ability.

According to the confusion matrix of the test set shown in Figure 4, emotion recognition for each category shows good performance: 94 samples in the anxiety category were correctly identified, 185 in the depression category, 302 in the anger category, 114 in the fear category, 202 in the sadness category, and 236 in the shame category. This indicates that the proposed multimodal fusion module and long-distance emotion fusion module play a role. The comprehensive capture of multimodal data enables the model to have a more accurate understanding of different emotions, and the long-distance emotion processing mechanism compensates for the interference of spatiotemporal differences in emotion recognition, resulting in high correct recognition numbers for each emotion category, especially showing outstanding performance in identifying emotions such as anger and shame.

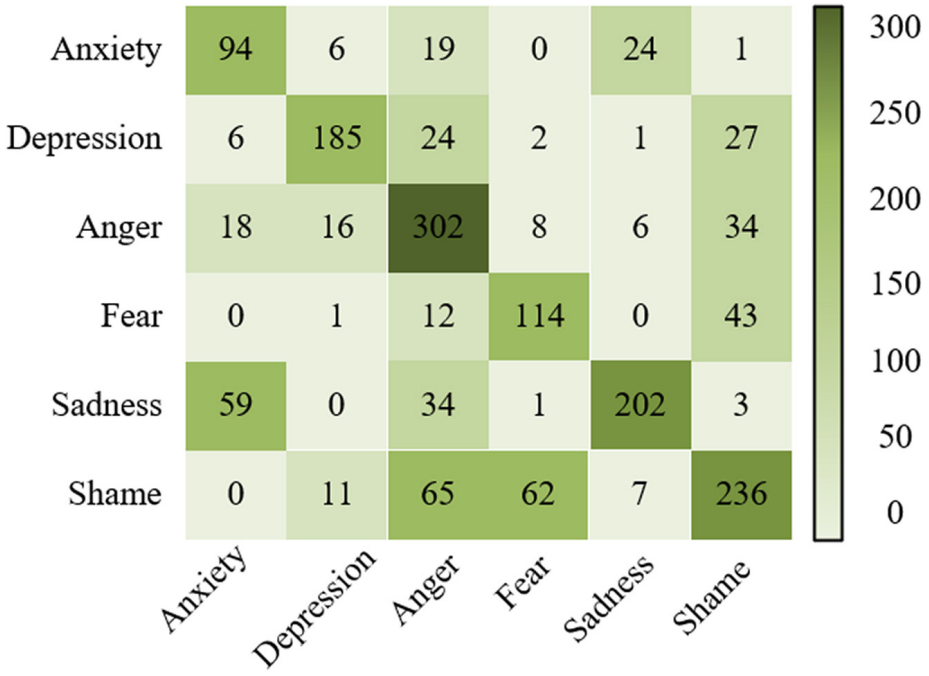


Fig. 4. Confusion matrix of the test set

Table 2. Unimodal emotion feature recognition results

Modality	Feature Extraction Method	Acc-2/%	F1/%	MAE	Corr
Chinese Text Semantics	GloVe	72.6	71.5	1.248	0.589
	FastText	74.5	72.6	1.236	0.623
Voice Prosody	LPCC	62.3	61.5	1.356	0.485
	PLP	63.8	63.4	1.389	0.521
	GFCC	71.5	71.8	1.342	0.513
Facial Micro-expression	HOG	56.8	55.9	1.562	0.468
	ORB	63.4	62.5	1.485	0.512
Multimodal	Proposed Model	83.5	83.9	0.756	0.779

According to Table 2, the classification results of unimodal emotion features show that the unimodal feature extraction methods such as Chinese text semantics (GloVe, FastText), voice prosody (LPCC, PLP, GFCC), and facial micro-expression (HOG, ORB) have limitations in indicators like Acc, F1, MAE, and Corr. For example, GloVe for Chinese text semantics has an Acc of 72.6%, LPCC for voice prosody has an Acc of 62.3%, and HOG for facial micro-expression has an Acc of 56.8%. However, the proposed multimodal model, which integrates three core modalities: text semantics, voice prosody, and facial micro-expression, achieves significantly better results with an Acc of 83.5%, an F1 of 83.9%, a lower MAE (0.756), and a higher Corr (0.779). This indicates that the multimodal fusion module compensates for the shortcomings of unimodal emotion information representation by comprehensively capturing multi-source data. Combined with the long-distance emotion fusion module's processing of spatiotemporal context and cross-regional emotional features, it verifies that the mobile network-based emotion analysis method for psychological health service dialogues can effectively improve emotion recognition performance.

Table 3. Multimodal emotion recognition results on test set

Model	Acc-2/%	F1/%	MAE	Corr
FER	81.2	78.6	0.915	0.678
Affectiva	83.5	81.5	0.856	0.682
POMS	81.8	81.6	0.814	0.735
PANAS	82.6	82.5	0.768	0.745
LIWC	82.5	82.4	0.758	0.736
Proposed Model	82.7	83.9	0.751	0.779

From Table 3, regarding the multimodal emotion classification results on the test set, compared with models such as FER, Affectiva, POMS, PANAS, and LIWC, the proposed model shows advantages in key indicators: Acc reaches 82.7%, although there are slight differences compared to some models, the F1 value leads at 83.9%, MAE (0.751) is lower than most comparison models, and Corr (0.779) is also at a higher level. This indicates that the designed multimodal fusion module effectively integrates multi-source data such as text, voice, and facial expressions, comprehensively capturing emotional information. The long-distance emotion fusion module

considers the spatiotemporal context and constructs cross-regional emotional feature mappings, optimizing emotion analysis performance. Even though the Acc indicator is not absolutely optimal, considering F1, MAE, and Corr, it still reflects the method's effectiveness in emotion recognition accuracy, error control, and feature correlation mining.

Based on the multimodal emotion recognition results on the mobile computing platform, psychological health interventions can be further promoted from the following strategies: After accurately identifying emotions such as anxiety and depression through multimodal fusion, real-time feedback and warnings can be triggered. Once abnormal emotional fluctuations are detected, such as long-term high anxiety features identified by the model, system warnings are immediately triggered and personalized intervention plans pushed; layered interventions: for mild emotional issues, push self-help resources such as breathing training and cognitive behavioral therapy dialogue guidance to regulate negative thoughts; for moderate to severe cases, automatically connect to human psychological consultation, using multimodal emotional data to assist consultants in making rapid judgments; Dynamic tracking and adjustment: long-term recording of emotional data to form visual tracking, analyze emotional triggering factors, iteratively optimize intervention strategies based on patterns, and simultaneously integrate long-distance emotional information such as spatiotemporal context to optimize the adaptability of cross-scenario and cross-regional interventions. With the help of the multimodal model, continuously improve the accuracy of intervention, building a closed-loop psychological health service from recognition to intervention and then to long-term management.

4 CONCLUSION

This study revolves around the "Emotion Recognition-Based Psychological Health Intervention Strategy on Mobile Computing Platform," focusing on emotion analysis of mobile network psychological health service dialogues. By designing multimodal fusion and long-distance emotion fusion modules, a precise emotion recognition model is constructed to support psychological health interventions. In the study, the multimodal fusion module integrates multi-source data such as text semantics, voice prosody, and facial micro-expressions. The long-distance emotion fusion module incorporates spatiotemporal context and cross-regional feature mappings. Experimental verification shows that in the multimodal emotion classification on the test set, the proposed model achieves an F1 value of 83.9%, and the indicators Corr (0.779) and MAE (0.751) outperform comparison models such as FER and LIWC. Ablation experiments and t-SNE visualization also show that collaborative multi-module design can improve emotional feature separability and recognition accuracy. From the perspective of technological innovation, this approach breaks through the limitations of unimodal recognition, and the multi-module fusion mechanism provides a more comprehensive and accurate solution for emotion recognition in mobile scenarios. From the perspective of application value, precise emotion recognition lays a solid foundation for psychological health interventions, supporting real-time warnings, layered intervention, and dynamic tracking. It fits the convenience of mobile platforms and helps build a "recognition-intervention-management" closed-loop service, which is of great significance in expanding the coverage of psychological health services and improving intervention efficiency.

This study has two limitations. First is the data and scenario adaptability: environmental noise in mobile scenarios and user privacy concerns may affect the quality and scale of multimodal data collection. Second is the depth of intervention

strategy implementation: currently, only the effectiveness of the emotion recognition model has been verified, and the full-process closure of manual collaboration and precise resource delivery in actual intervention is yet to be fully verified. In future research directions, data ecosystem construction can be deepened, and privacy computing technologies can be explored to expand multimodal and cross-regional datasets under the premise of ensuring data security. The closed-loop verification of intervention strategies can be promoted by testing the full process of "recognition-intervention-feedback" in real psychological consultation scenarios and optimizing the resource delivery logic. The generalization boundary of the model can also be expanded by studying the adaptability of multimodal fusion in special groups and combining meta-learning and adaptive technologies to improve model robustness, promoting the deep application of mobile platform psychological health services from "accurate recognition" to "effective intervention."

5 REFERENCES

[1] G. Chen, "Design and application of scenario-based perception of smart wearable device interaction method," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 18, no. 13, pp. 69–81, 2024. https://doi.org/10.3991/ijim.v18i13.49071

[2] F. Yang, "Application of mobile technology in enterprise management: From mobile office to intelligent decision support," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 18, no. 18, pp. 4–18, 2024. https://doi.org/10.3991/ijim.v18i18.51493

[3] A. J. Mills, R. T. Watson, L. Pitt, and J. Kietzmann, "Wearing safe: Physical and informational security in the age of the wearable device," *Business Horizons*, vol. 59, no. 6, pp. 615–622, 2016. https://doi.org/10.1016/j.bushor.2016.08.003

[4] M. H. Ryu, J. Kim, and S. Kim, "Factors affecting application developers' loyalty to mobile platforms," *Computers in Human Behavior*, vol. 40, pp. 78–85, 2014. https://doi.org/10.1016/j.chb.2014.08.001

[5] M. Ciman and O. Gaggi, "An empirical analysis of energy consumption of cross-platform frameworks for mobile development," *Pervasive and Mobile Computing*, vol. 39, pp. 214–230, 2017. https://doi.org/10.1016/j.pmcj.2016.10.004

[6] K. S. Staykova and J. Damsgaard, "Adoption of mobile payment platforms: Managing reach and range," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 11, no. 3, pp. 65–84, 2016. https://doi.org/10.4067/S0718-18762016000300006

[7] L. Mu, B. Du, and X. Hou, "A study on the improvement of college students' psychological pressure and anxiety by using English psychological script activities," *Frontiers in Psychology*, vol. 13, p. 878479, 2022. https://doi.org/10.3389/fpsyg.2022.878479

[8] C. Johnson and J. Taylor, "More than bullshit: Trash talk and other psychological tests of sporting excellence," *Sport, Ethics and Philosophy*, vol. 14, no. 1, pp. 47–61, 2020. https://doi.org/10.1080/17511321.2018.1535521

[9] G. Orrù, R. Ciacchini, A. Gemignani, and C. Conversano, "Psychological intervention measures during the COVID-19 pandemic," *Clinical Neuropsychiatry*, vol. 17, no. 2, pp. 76–79, 2020. https://doi.org/10.36131/CN20200208

[10] K. Hämmerli, H. Znoj, and J. Barth, "The efficacy of psychological interventions for infertile patients: A meta-analysis examining mental health and pregnancy rate," *Human Reproduction Update*, vol. 15, no. 3, pp. 279–295, 2009. https://doi.org/10.1093/humupd/dmp002

[11] E. Melguizo-Ibáñez, J. Cachón-Zagalaz, G. González-Valero, P. Puertas-Molero, L. García-Pérez, and J. L. Ubago-Jiménez, "Emotional status and psychological well-being in the educational opposition process," *Social Sciences*, vol. 12, no. 12, p. 685, 2023. https://doi.org/10.3390/socsci12120685

[12] J. M. Flujas-Contreras, A. García-Palacios, and I. Gómez, "Effectiveness of a web-based intervention on parental psychological flexibility and emotion regulation: A pilot open trial," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, p. 2958, 2021. https://doi.org/10.3390/ijerph18062958

[13] S. K. Khare and V. Bajaj, "An evolutionary optimized variational mode decomposition for emotion recognition," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 2035–2042, 2020. https://doi.org/10.1109/JSEN.2020.3020915

[14] S. Taran and V. Bajaj, "Emotion recognition from single-channel EEG signals using a two-stage correlation and instantaneous frequency-based filtering method," *Computer Methods and Programs in Biomedicine*, vol. 173, pp. 157–165, 2019. https://doi.org/10.1016/j.cmpb.2019.03.015

[15] F. Yan, Z. Guo, A. M. Iliyasu, and K. Hirota, "Multi-branch convolutional neural network with cross-attention mechanism for emotion recognition," *Scientific Reports*, vol. 15, 2025. https://doi.org/10.1038/s41598-025-88248-1

[16] A. Nalwaya, K. Das, and R. B. Pachori, "An automated framework for human emotion detection from multichannel EEG signals," *IEEE Sensors Journal*, vol. 24, no. 13, pp. 20920–20927, 2024. https://doi.org/10.1109/JSEN.2024.3398050

[17] I. Hosseini, M. Z. Hossain, Y. Zhang, and S. Rahman, "Deep learning model for simultaneous recognition of quantitative and qualitative emotion using visual and bio-sensing data," *Computer Vision and Image Understanding*, vol. 248, p. 104121, 2024. https://doi.org/10.1016/j.cviu.2024.104121

[18] C. Xu, P. Du, Z. Feng, Z. Meng, T. Cao, and C. Dong, "Multi-modal emotion recognition fusing video and audio," *Applied Mathematics & Information Sciences*, vol. 7, no. 2, pp. 455–462, 2013. https://doi.org/10.12785/amis/070205

[19] J. Zhang, L. Xing, Z. Tan, H. Wang, and K. Wang, "Multi-head attention fusion networks for multi-modal speech emotion recognition," *Computers & Industrial Engineering*, vol. 168, p. 108078, 2022. https://doi.org/10.1016/j.cie.2022.108078

6 AUTHORS

Yang Li received the B.S. degrees at the Hebei University of Economics and Business, Shijiazhuang, China, in 2003, and the M.S. degree in Psychology from Hebei University, Baoding, China, in 2006. He is currently pursuing the Ph.D. degree at Beijing Normal University, Beijing, China. His research interests include mental health, psychological crisis intervention and psychological evaluation (E-mail: 202132061004@mail.bnu.edu.cn).

Lihua Peng received the B.S. degrees at the Hebei University, Baoding, China, in 1999, and the M.S. degree in Psychology from Hebei University, Baoding, China, in 2004. Her research interests include mental health, social support and social integration (E-mail: penglihuamail@126.com).