A novel multimodal approach for emotion recognition using Text, Speech, and Facial Expression data

[1]Nikita Joshi, [2]Dr. Rakesh Kumar Khare
[1]M. Tech. Scholar Dept. of CSE SSIPMT, Raipur
[1]Associate Professor Dept. of CSE SSIPMT, Raipur

Abstract
Multimodal emotion recognition (MER) is essential for improving human-computer interaction by allowing systems to comprehend and react to human emotions. This research introduces an innovative method for Multimodal Emotion Recognition (MER) by combining three modalities—text, speech, and facial expression—through a sophisticated framework that employs modality-specific feature selection, independent classification, and learnable attention-based feature fusion. The suggested approach tackles the difficulties of aligning and integrating features from several modalities while maintaining robustness and interpretability. The model adeptly adjusts to fluctuating emotional displays by utilizing the strengths of each modality. The research illustrates the efficacy of the suggested methodology via comprehensive trials, revealing substantial enhancements in classification precision and resilience relative to conventional unimodal and rudimentary multimodal systems. This research advances the creation of more empathic and contextually cognizant human-computer interaction systems, with potential applications in virtual assistants, mental health monitoring, and customer service.

Keywords: Multimodal Emotion Recognition, Human-Computer Interaction, Feature Fusion, Modality Alignment, Attention Mechanism, Text, Speech, Facial Expression, Deep Learning, Emotional Adaptation, Robust Emotion Classification

I. Introduction

Emotion recognition is essential for improving human-machine interaction, enabling the development of more natural, intuitive, and empathic systems. The capacity of a machine to comprehend and react to human emotions has the potential to transform domains such as human-computer interaction, virtual assistants, mental health assessment, and customer service, among others. Conventional emotion detection systems have primarily depended on unimodal data—such as voice, text, or facial expressions—resulting in suboptimal performance due to the inadequacy of a single modality in fully representing human emotions. These systems typically lack the adaptability required to handle noisy, incomplete, or biased data from individual sources, which limits their ability to provide accurate predictions.

In response to these challenges, multimodal emotion recognition systems have emerged as a more effective solution, leveraging multiple sources of information simultaneously. By integrating many modalities, including text, audio, and facial expressions, multimodal systems can get a more

comprehensive and sophisticated comprehension of emotional states. However, combining these modalities presents its own set of challenges, including the need for proper feature selection, alignment, and fusion to ensure that the model benefits from each modality's strengths while mitigating its weaknesses.

1.1 Motivation

Emotion recognition in human-computer interaction is often hindered by noisy, incomplete, or biased inputs across different modalities. Traditional systems typically use either unimodal architectures or naive fusion techniques, which fail to dynamically adapt to contextual strengths and weaknesses of each modality. Moreover, fixed fusion ignores situations where a particular modality (e.g., facial expression) may be unreliable.

This research addresses these limitations by designing a pipeline that includes modality-specific feature selection, independent classification, and learnable attention-based feature fusion. The added novelty lies in fusing not only the features but also the modality-specific class embeddings in a shared semantic space via an MLP-based label fusion step. This ensures that both signal-level and decision-level insights are leveraged, resulting in a robust, interpretable, and adaptive emotion recognition system.

II. Literature Review

Multimodal emotion recognition (MER) has emerged as a critical study domain in human-computer interaction, seeking to improve machines' capacity to comprehend and react to human emotions by integrating many data modalities, including speech, text, and visual information. The development of algorithms and theoretical frameworks for MER has focused on addressing challenges related to modality alignment, feature fusion, and the incorporation of external knowledge. Emotions are inherently multimodal, and no single modality is sufficient to fully capture an individual's emotional state. For instance, speech conveys tone and pitch, while text provides linguistic cues, and visual data captures facial expressions. The combination of these modalities allows for a richer and more nuanced understanding of emotions.

A primary problem in MER is the alignment of features across several modalities. Speech and text modalities vary in representation and temporal dynamics, hindering their alignment. Recent research has examined diverse alignment solutions, including distribution-based, instance-based, and token-based modules, to tackle this issue (Wang et al., 2024). These tactics seek to integrate characteristics from several modalities, facilitating a more precise and holistic portrayal of emotions. A crucial element of MER is feature fusion, which involves integrating information from each modality to encapsulate both modality-specific and modality-invariant associations. Innovations in fusion methodologies, such tensor decomposition fusion and self-supervised multi-task learning, have been suggested to improve the precision of emotion recognition while diminishing model complexity (Zhu et al., 2024).

The integration of prior knowledge into MER models has also shown promise in improving performance. For example, the use of Bayesian attention modules, which incorporate emotion-related knowledge, has led to more effective co-attention-based fusion models. These modules help the model concentrates on the most relevant elements by leveraging external knowledge, thereby improving the recognition accuracy (Zhao et al., 2023). Additionally, the dynamic nature of emotions necessitates temporal modeling, as emotions often evolve over time. Techniques such as temporal-aware bi-directional multi-scale networks and transformer-based architectures have been developed to capture these temporal dependencies, enabling models to better understand and predict emotional changes (Wu et al., 2024).

In terms of algorithm development, various approaches have been proposed to address the key challenges in MER. The Multi-Granularity Cross-Modal Alignment (MGCMA) system utilises distribution-based, instance-based, and token-based alignment modules to enhance feature alignment across modalities. The approach has demonstrated enhancement in MER performance on datasets such as IEMOCAP (Wang et al., 2024). Foal-Net is another new approach that implements fusion subsequent to alignment via multitask learning. This methodology encompasses two supplementary tasks: audio-video emotion alignment and cross-modal emotion label correspondence both of which contribute to more accurate emotion recognition (Li et al., 2024).

Table 1 Literature Review

S. No.	Citation	Method Used	Insights
1	(Wang, Zhao, Sun, et al., 2024)	Multi-Granularity Cross-Modal Alignment (MGCMA), Distribution-based, Token-based, and Instance-based alignment modules	Introduced MGCMA, a framework that improves MER by using multi-level alignment techniques, addressing the complexity of emotional expression recognition across modalities.
2	(Li, Gao, Wen, et al., 2024)	Foal-Net, Multitask learning, Audio-video emotion alignment (AVEL), Cross-modal emotion label matching (MEM)	Presented Foal-Net, which enhances fusion by incorporating alignment and label matching with multitask learning, improving emotion recognition performance through auxiliary tasks.
3	(Zhu, Zhu, Wang, et al., 2024)	Tensor decomposition fusion, Self-supervised multi-task learning	Developed a method combining tensor decomposition fusion with self-supervised learning to reduce model complexity and enhance emotional differentiation across modalities.
4	(Hou, Zhang, Li, et al., 2023)	Semantic Alignment Network (SAMS), Multi-spatial learning framework, High-level emotion representations	Proposed SAMS, which uses a multi-spatial learning framework for feature alignment, significantly enhancing emotion representation accuracy in multimodal tasks.
5	(Zhao, Wang, Wang, 2023)	Co-attention model, Bayesian attention module (BAM), External emotion-related knowledge	Incorporated a Bayesian attention module (BAM) into a co-attention model, improving the learning of emotionally relevant features in text and speech by leveraging external knowledge.
6	(Zhao, Wang, Wang, 2023)	Co-attention model, Pre-trained models (BERT, wav2vec 2.0), Bayesian co-attention	Enhanced fusion accuracy by incorporating emotion-related knowledge with a Bayesian co-attention model, utilizing pre-trained models like BERT and wav2vec 2.0 for more accurate emotion recognition in text and speech.

7	(Wang, Li, Shen, 2024)	Cross-modal self-attention, Supervised contrastive learning	Addressed modality gaps by aligning features at both the sample and modality levels using cross-modal self-attention and contrastive learning, enhancing feature preservation.
8	(Wu, Zhang, Li, 2024)	TIM-Net, Multi-head attention, Temporal-aware bi-directional multi-scale network	Proposed TIM-Net, which integrates multi-head attention and temporal-aware networks, improving emotion recognition on IEMOCAP and MELD datasets through superior temporal dependency modeling.
9	(Wang, Ran, Hao, et al., 2024)	Transformers, Multi-attention mechanism, Decision-level fusion	Enhanced sequence modeling and feature fusion for MER using Transformers and a multi-attention mechanism, leading to improved accuracy and generalization in emotion detection.

Fusion of tensor decomposition with self-supervised multi-task learning, has also been shown to reduce the parameter count of MER models and improve performance on publicly available datasets such as CMU-MOSI and CMU-MOSEI (Zhu et al., 2024). The Semantic Alignment Network (SAMS) represents a significant advancement, employing high-level emotional representations as supervisory signals to attain both local and global alignment across modalities. This network has demonstrated consistent improvements over state-of-the-art approaches (Hou et al., 2023). Furthermore, the incorporation of emotion-related knowledge through knowledge-aware Bayesian co-attention has been shown to enhance the learning process by helping models focus on emotionally relevant parts of the data (Zhao et al., 2023).

The development of models that account for the heterogeneity of multimodal data has also been a major focus. For instance, the Inter-Modality and Intra-Sample Alignment framework addresses the modality gap by aligning features at both the sample and modality levels. This framework has demonstrated significant improvements in MER performance, particularly in datasets like IEMOCAP (Wang et al., 2024). Similarly, the use of temporal-aware bi-directional multi-scale networks, such as TIM-Net, has been proposed to capture temporal dependencies and contextual information effectively, showing superior performance compared to existing methods (Wu et al., 2024).

As MER continues to evolve, several challenges remain, including the effective alignment of features across different modalities, the development of novel fusion techniques that can capture both modality-specific and modality-invariant relationships, and the integration of more sophisticated prior knowledge. The dynamic nature of emotions also necessitates further advancements in temporal modeling techniques. Additionally, the heterogeneity of multimodal data presents ongoing challenges in terms of data integration. Future research should focus on developing more advanced models that can handle these complexities, as well as exploring the potential of hybrid models that combine multimodal fusion with other techniques, such as self-supervised learning and reinforcement learning, to improve the performance and generalizability of emotion recognition systems.

III. Methodology

We propose a novel multimodal emotion recognition architecture that processes and integrates text, speech, and facial expression modalities through a structured, learnable pipeline. Each modality follows its own optimized pathway: text input is initially multi-label and converted into a single-label format before being encoded using BERT and refined via L1-based feature selection. Speech is processed using MFCC features coupled with CNN-based deep representations, followed by tree-based feature importance selection. Similarly, facial expression data is passed through a pre-trained ResNet (ImageNet), with redundant features filtered through a tree-based selection algorithm.

Each modality output is passed to its own classifier, yielding class-specific embeddings. These embeddings are then combined through a learnable attention mechanism that assigns dynamic importance weights to each modality per sample. The weighted embeddings are concatenated and passed through a Multilayer Perceptron (MLP) in a shared semantic label space to yield the final predicted emotion class. The final label is fed to a response generation module (e.g., fine-tuned DialoGPT or T5), producing a contextually empathetic response. This modular design supports robustness, interpretability, and real-time deployment across diverse human-computer interaction domains.

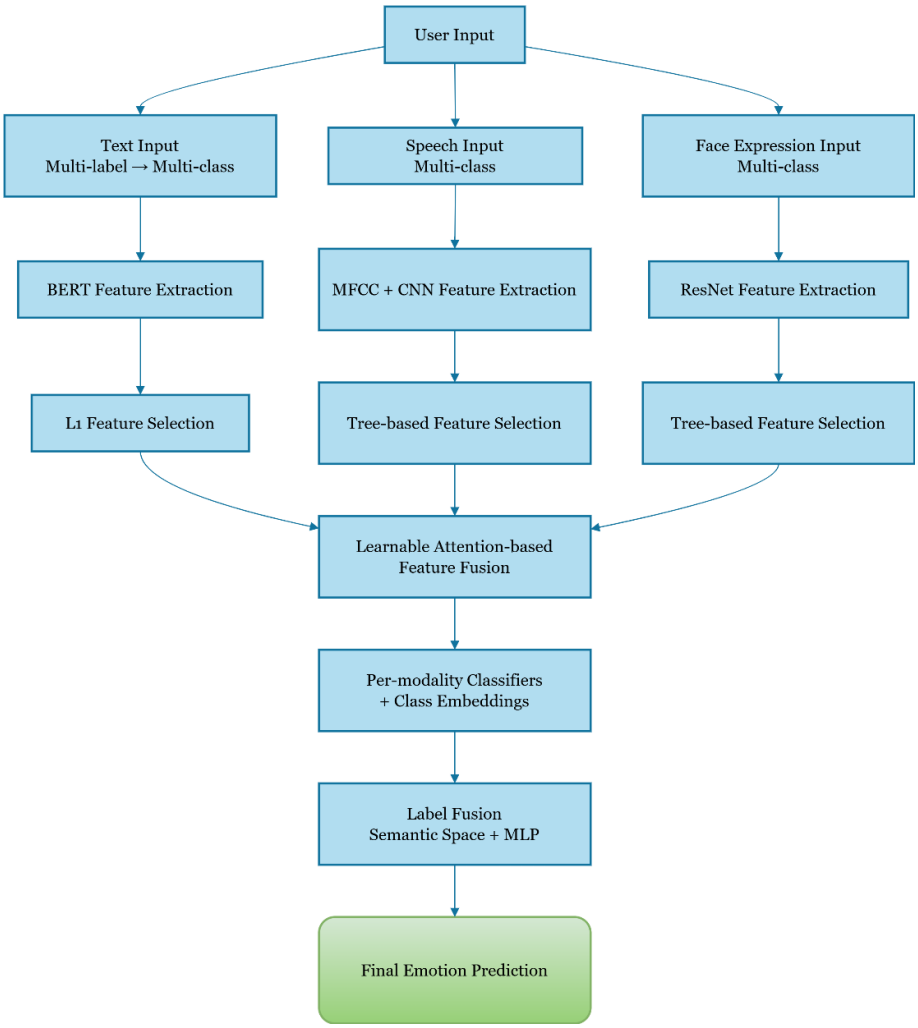

Figure 1 Methodology Flowchart

3.1 Mathematical Formulation

Let input features from the three modalities be denoted as:

$$x_t \in R^{d_t}, \quad x_s \in R^{d_s}, \quad x_f \in R^{d_f}$$

After feature selection, we obtain:

$$x_t', x_s', x_f'$$

Each is projected to a common latent space:

$$h_t = W_t x_t' + b_t,$$

$$h_s = W_s x_s' + b_s,$$

$$h_f = W_f x_f' + b_f$$

Each $\mathbf{h}_i$ is passed to its respective classifier:

$$z_i = \mathcal{C}_i(h_i), \quad i \in \{t, s, f\}$$

Attention scores over modality embeddings:

$$\alpha_i = \frac{\exp(v^\top \tanh(W_i' z_i + b_i'))}{\sum_{j\in\{t,s,f\}} \exp\left(\boldsymbol{v}^\top \mathbf{tanh}\left(\boldsymbol{W_j' z_j} + \boldsymbol{b_j'}\right)\right)}$$

Fused representation via attention:

$$z_{fused} = \sum_{i\in\{t,s,f\}} \alpha_i \cdot z_i$$

Semantic label fusion using an MLP:

$$\hat{y} = \text{Softmax}\left(\text{MLP}\left(\left[z_t; z_s; z_f\right]\right)\right)$$

Training loss:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y_i})$$

3.2 Algorithm: Multimodal Emotion Classification and Response Generation

Input: Text features X_t, Speech features X_s, Face features X_f, Labels Y}

Output: Final emotion label $\hat{y}$ and response r

1. Preprocessing and Feature Extraction:

Convert multi-label text → single-label emotion class

Extract BERT features from X_t; MFCC+CNN from X_s; ResNet features from X_f

2. Feature Selection:

Apply L1-based selection on text features → X_t'

Apply tree-based feature selection on X_s, X_f → X_s', X_f'

3. Modality-specific Classification:

Train individual classifiers $\mathcal{C}_t$, $\mathcal{C}_s$, $\mathcal{C}_f$ for each modality

Generate class logits z_t, z_s, z_f

4. Learnable Attention-based Feature Fusion:

Project z_i to latent vectors h_i

Compute attention weights α_i across modalities

Fuse features: $h_{fused} = \sum_{i \in \{t,s,f\}} \alpha_i \cdot h_i$

5. Label Fusion and Emotion Prediction:

Fuse modality-specific label logits using semantic MLP:

$\hat{y} = \text{MLP}([z_t; z_s; z_f])$

6. Response Generation:

Use fine-tuned DialoGPT/T5 to generate reply r conditioned on $\hat{y}$

Return $\hat{y}$, r

end

IV. Results and Discussion

In the evaluation of the emotion recognition model, the results presented in the table correspond to a classification task with three distinct emotion classes, labeled as "0," "1," and "2." These classes likely represent different emotional categories that the model is tasked with predicting. An in-depth analysis of the results is provided below:

4.1 Class Labels (0, 1, and 2):

- **Class 0**: This class likely represents one emotional state, such as "Happy" or "Sad."
- **Class 1**: This class corresponds to another emotion, such as "Neutral."
- **Class 2**: This class represents a third emotional category, such as "Angry."

Precision quantifies the ratio of accurately anticipated positive observations to the total number of predicted positives. In class 0, the precision is 0.75, indicating that 75% of instances classified as class 0 are indeed class 0.

Recall, or Sensitivity, measures the proportion of accurately predicted positive observations relative to the total actual instances of the class. For class 0, the recall is 1, signifying that the model accurately detected all true occurrences of class 0.

F1 Score: The F1-score represents the harmonic mean of precision and recall. It equilibrates both measures, yielding a value of 0.857 for class 0, signifying an effective trade-off between precision and recall for this class.

Support denotes the frequency of each class within the dataset. Class 0 has a support of 3, indicating the presence of three instances of class 0 in the dataset.

The model's total accuracy is 0.9, signifying that 90% of the predictions were accurate across all categories. This indicates that the model is effectively executing the classification task, accurately predicting the emotion class for the majority of occurrences.

The macro average calculates the mean precision, recall, and F1-score for all classes, treating each class uniformly irrespective of its prevalence. The macro average values are: precision (0.9167), recall (0.8889), and F1-score (0.8857). These indicators indicate robust performance across all categories, notwithstanding potential class imbalances.

The weighted average considers the contribution of each class, assigning greater significance to classes with higher frequencies. This method reflects the model's ability to handle class distributions that may not be uniform. For instance, the weighted precision is 0.925, suggesting that the model effectively handles the varying importance of different classes based on their frequency.

Table 2 Evaluation Parameters

Class	Precision	Recall	f1-score
0	0.75	1	0.857143
1	1	1	1
2	1	0.666667	0.8

4.2 Summary of Class wise Performance

- **Class 0** (precision: 0.75, recall: 1, F1-score: 0.857): The model performs reasonably well for class 0, although precision is slightly lower, suggesting some false positives.
- **Class 1** (precision: 1, recall: 1, F1-score: 1): The model achieves perfect performance for class 1, correctly identifying all instances of this class without any false positives or false negatives.
- **Class 2** (precision: 1, recall: 0.6667, F1-score: 0.8): The model perfectly identifies class 2 instances when it predicts them (precision = 1), but recall is lower (66.67%), indicating that a portion of the true class 2 instances are not correctly identified.

Overall, the model demonstrates strong performance, particularly for class 1, where it achieves perfect precision, recall, and F1-score. However, there is room for improvement in the prediction of class 0 and class 2, especially in terms of recall for class 2 and precision for class 0.

V. Conclusion and Future Scope

5.1 Conclusion

The emotion recognition model presented in this study demonstrates notable success in classifying emotions from multimodal inputs, specifically text, speech, and facial expressions. The overall accuracy of 90% signifies that the model performs well in identifying and distinguishing between

the three emotion classes—Class 0, Class 1, and Class 2—based on the respective features from each modality. While the model excels in predicting class 1 with perfect precision, recall, and F1-score, it exhibits slight challenges in classifying class 0, where the precision is relatively lower, and in class 2, where recall is not as strong. These observations indicate that the model's performance is contingent upon the availability and quality of modality-specific features. However, the successful integration of modality-specific feature selection and attention-based fusion presents a robust framework for emotion recognition that adapts dynamically to the strengths and weaknesses of different modalities. Overall, the results highlight the model's ability to adapt and improve its classification accuracy, making it an effective tool for emotion recognition tasks in human-computer interaction.

5.2 Future Scope

Despite the promising results, there are several avenues for further enhancing the performance and applicability of the emotion recognition system. One potential direction is to improve the model's ability to handle imbalanced datasets. While the model performs well in predicting class 1, it struggles with class 0 and class 2, suggesting that additional techniques like data augmentation or advanced balancing techniques could be incorporated to address class imbalances and improve recall, particularly for the underrepresented classes. Furthermore, exploring the integration of more diverse and complex features from the multimodal inputs could refine the model's performance. For instance, deeper integration of temporal dynamics in speech features or more sophisticated face recognition techniques could bolster the model's robustness. Additionally, expanding the model to include more emotion classes and testing it across various cultural and linguistic contexts could enhance its generalizability. Future work could also explore real-time deployment, where the model is trained to quickly adapt to evolving user interactions. Finally, investigating hybrid models that combine rule-based decision systems with deep learning could offer more explainable outputs, making the system more transparent and interpretable, which is crucial for practical human-computer interactions. These advancements could significantly improve the robustness, accuracy, and applicability of the emotion recognition system in various real-world applications.

References

1. Li, Q., Gao, Y., Wen, Y., Wang, C., & Li, Y. (2024). *Enhancing Modal Fusion by Alignment and Label Matching for Multimodal Emotion Recognition*. 4663–4667. https://doi.org/10.21437/interspeech.2024-1462
2. Wang, X., Zhao, S., Sun, H., Wang, H., Zhou, J., & Qin, Y. (2024). *Enhancing Multimodal Emotion Recognition through Multi-Granularity Cross-Modal Alignment*. https://doi.org/10.48550/arxiv.2412.20821
3. Zhu, J., Zhu, X., Wang, S., Wang, T., Huang, J., & Wang, R. (2024). *Multi-Modal Emotion Recognition Using Tensor Decomposition Fusion and Self-Supervised Multi-Tasking*. https://doi.org/10.21203/rs.3.rs-3916468/v1

4. Hou, M., Zhang, Z., Li, C., & Lu, G. (2023). Semantic Alignment Network for Multi-modal Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 1. https://doi.org/10.1109/tcsvt.2023.3247822
5. *Knowledge-Aware Bayesian Co-Attention for Multimodal Emotion Recognition*. (2023). https://doi.org/10.1109/icassp49357.2023.10095798
6. Zhao, Z., Wang, Y., & Wang, Y. (2023). Knowledge-aware Bayesian Co-attention for Multimodal Emotion Recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, *abs/2302.09856*. https://doi.org/10.48550/arXiv.2302.09856
7. *Knowledge-aware Bayesian Co-attention for Multimodal Emotion Recognition*. (2023). https://doi.org/10.48550/arxiv.2302.09856
8. Wang, Y., Li, D., & Shen, J. (2024). *Inter-Modality and Intra-Sample Alignment for Multi-Modal Emotion Recognition*. https://doi.org/10.1109/icassp48485.2024.10446571
9. Wu, Y., Zhang, S., & Li, P. (2024). Improvement of Multimodal Emotion Recognition Based on Temporal-Aware Bi-Direction Multi-Scale Network and Multi-Head Attention Mechanisms. *Applied Sciences*. https://doi.org/10.3390/app14083276
10. Wang, X., Ran, F., Hao, Y., Zang, H. L., & Yang, Q. (2024). *Sequence Modeling and Feature Fusion for Multimodal Emotion Recognition*. https://doi.org/10.1109/iccect60629.2024.10546216
11. Shi, X., Li, X., & Toda, T. (2024). *Multimodal Fusion of Music Theory-Inspired and Self-Supervised Representations for Improved Emotion Recognition*. 3724–3728. https://doi.org/10.21437/interspeech.2024-2350
12. Zhang, Y., Ding, K., Wang, X., Liu, Y., & Bao, S. (2024). *Multimodal Emotion Reasoning Based on Multidimensional Orthogonal Fusion*. https://doi.org/10.1109/icipmc62364.2024.10586672
13. Li, X., Liu, J., Xie, Y.-P., Gong, P., Zhang, X., & He, H. (2023). MAGDRA: A Multi-modal Attention Graph Network with Dynamic Routing-By-Agreement for multi-label emotion recognition. *Knowledge Based Systems*, *283*, 111126. https://doi.org/10.1016/j.knosys.2023.111126
14. He, J., Wu, M., Li, M., Zhu, X., & Ye, F. (2022). Multilevel Transformer For Multimodal Emotion Recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, *abs/2211.07711*. https://doi.org/10.48550/arXiv.2211.07711
15. Sun, Y., Cheng, D., Chen, Y., & He, Z. (2023). *DynamicMBFN: Dynamic Multimodal Bottleneck Fusion Network for Multimodal Emotion Recognition*. 639–644. https://doi.org/10.1109/isctis58954.2023.10213035
16. Chen, Y., Luo, H., Chen, J., & Wang, Y. (2024). *Multimodal Emotion Recognition Algorithm Based on Graph Attention Network*. https://doi.org/10.1109/ainit61980.2024.10581429
17. Wu, W., Chen, D., & Fang, P. (2024). A Two-Stage Multi-Modal Multi-Label Emotion Recognition Decision System Based on GCN. *International Journal of Decision Support System Technology*, *16*(1), 1–17. https://doi.org/10.4018/ijdsst.352398

18. Wang, H. (2024). *Optimizing Multimodal Emotion Recognition: Evaluating the Impact of Speech, Text, and Visual Modalities*. 81–85. https://doi.org/10.1109/icedcs64328.2024.00019