Detection and Analysis of Human Mental State Based on Multimodal Information

Jiahao Luo [1], Zicheng Wang [2, *]

[1] School of Electrical and Electronic Information Engineering, Hubei Polytechnic University, Huangshi, China

[2] Portland Institute, Nanjing University of Posts and Telecommunications, Nanjing, China

* Corresponding Author Email: P22000109@njupt.edu.cn

Abstract. With the continuous evolution and refinement of intelligent sensing technologies, the development and application of multimodal intelligent information in detecting human mental states have become a key direction in the advancement of intelligent technologies, surpassing traditional unimodal detection methods. Multimodal detection technology that integrates multiple information sources to reflect human mental states has been widely applied in fields such as medical care and interrogation. However, existing detection methods generally have disadvantages such as low precision and detection equipment being susceptible to environmental influences. Based on this background, this paper explores a framework for detecting and analyzing human mental states based on multimodal information. This framework integrates multi-source data, including non-contact sensing, contact-based physiological signal acquisition, and emerging dataset construction. This framework combines feature level, decision-making-level level and end-to-end multimodal fusion methods to achieve the deep synergy of cross-modal features. It enables the coordinated utilization of multiple human signal technologies, establishing a comprehensive data analysis framework for multimodal feature fusion. The rational use of this framework can significantly improve the robustness of the detection method. This study proposes innovative improvement solutions for applications in scenarios such as medical monitoring and judicial interrogations.

Keywords: Human mental state, multimodal information, framework.

1. Introduction

With the rapid advancement of intelligent technologies, the demand for intelligent perception in complex scenarios is becoming increasingly urgent. The detection of human mental states, as a vital branch of intelligent perception, holds significant application value in fields such as medical monitoring and judicial interrogation.

Globally, the World Health Organization (WHO) has designated mental health as a core indicator of the 2030 Agenda for Sustainable Development [1]; According to the 2023 Global Wellness Economy Monitor report, annual investments in real-time mental state monitoring across public safety and digital healthcare sectors in North America and the European Union alone have surpassed $26 billion [2]. Against this macro backdrop, “the detection and analysis of human mental states” has emerged as a critical research topic spanning multiple fields, including medical monitoring and judicial interrogation.

Traditional single-modal approaches to assessing human mental states typically rely on subjective questionnaire reports, one-way cameras, or single physiological sensors for mental state evaluation. Subjective questionnaire reports are often susceptible to biases stemming from cultural differences and social expectations, leading to emotional states reported in the questionnaires that do not align with reality [3]; Single cameras often exhibit lower accuracy under varying environmental conditions, such as changes in lighting, obstructions, or makeup [4]; Moreover, single physiological sensors such as electroencephalogram (EEG) and heart rate monitoring are often affected by motion artifacts, wireless interference, and individual variations [5]. Therefore, the accuracy and robustness of traditional single-mode detection methods are insufficient to meet requirements.

As scientific research continues to advance, the emergence of multimodal intelligent perception has provided new approaches for detecting mental states. By integrating multi-source data, including voice, facial expressions, electroencephalogram signals, skin conductance, heart rate variability, and radio frequency signals, researchers can characterize an individual's state across physiological, behavioral, and cognitive dimensions, thereby overcoming the limitations of single-modal detection in terms of accuracy, robustness, and interpretability. In interrogation scenarios, multimodal approaches can reveal deception by integrating the interplay of voice-facial inconsistencies, abnormal brainwave patterns, and skin conductance responses.; In medical settings, multimodal data fusion can be utilized for screening and assessing psychological disorders such as depression and anxiety, providing objective evidence for clinical interventions.

To address these requirements, this study first established a multimodal data acquisition framework integrating non-contact sensing (RF, infrared, visual), contact-based physiological signals (EEG, GSR, BCG, heart rate), and emerging datasets (laboratory simulations and real-world scenarios) to enable cross-scenario, cross-modal acquisition of mental state data. Second, at the feature fusion level, this study systematically explored multiple approaches spanning feature-level (combining MFCC with EEG P300), decision-level (weighted voting and attention mechanisms), and end-to-end fusion (lightweight UT-Adapter and interpretable ERNN). It compared the strengths and weaknesses of different strategies in terms of accuracy and robustness. Subsequently, this study validates the proposed method in typical tasks such as deception detection, mental health assessment, acute stress monitoring and emotion recognition, and compares the performance of the proposed method with that of traditional multimodal methods, demonstrating the significant superiority of multimodal methods.

2. Multimodal Data Acquisition Technology and Multimodal Feature Fusion Method

2.1. Multimodal Data Acquisition Technology

At present, the data acquisition technology of multimodal mental detection mainly focuses on physiological signals, behavioral signals and neuroimaging. In the field of physiological signal acquisition, the six-channel synchronous acquisition system (EEG/GSR/BCG/temperature) developed by Qin, S. M. presents a relatively complete hardware design. Its innovation is reflected in the dynamic gain adjustment (0.128-176 times) of programmable instrument amplifier (PGA281), common mode noise suppression technology (right leg drive+shielding) and sensor interface optimization (dry/wet electrodes are used in different scenarios). The system is miniaturized through a four-layer PCB split design (the acquisition front end is only 9.2 × 5.0mm), and the noise control reaches EEG channel<3μV rms [6] [7]. Ernst, H. et al. proposed a three-system coordination scheme, AD Instruments to collect PPG/ECG/EDA, task force monitor to monitor continuous blood pressure, IDS camera to capture facial video (100fps), and realize cross-device millisecond alignment through the master clock (power lab 16/35) and synchronous waveform [8] [9]. For behavior signal acquisition, Jiang, Z. et al. achieved HIPAA-compliant remote video interviews through the zoom platform. Synchronously recorded audio and video streams, and extracted cardiovascular signals from facial ROI using rPPG technology. Guo, X. et al. creatively adopted the reality show game scene (dolos dataset), collected 1675 video clips in the deception situation with strong motivation, and labeled 25 types of facial action units through the MUMIN coding scheme [10]. Neuroimaging acquisition is reflected in Liu, R. et al.'s article, integrating functional magnetic resonance (fMRI) and structural magnetic resonance (sMRI), using C-PAC/DPARSF standardized pretreatment pipeline to extract functional connectivity (CC200 map divided into 200 ROI) and gray matter volume characteristics, covering the diagnosis of autism, ADHD and other brain diseases [11-15]. It is worth noting that data collection is exhibiting an imperceptible trend. The cloud dialogue agent system (Tina) in Cohen, J. et al. automatically guides the interview process and monitors the voice duration in real time [11]. Jiang, Z. et al. continuously collected 13.85 ± 6.94 days' motion data through wearable devices (wrist accelerometers) [16].

2.2. Evolution of Multimodal Feature Fusion Methods

The feature fusion method is gradually deepened from early splicing, attention mechanism, to decoupling learning. Early fusion is mainly based on simple feature-level splicing. For example, Middleya, A. I. et al. directly connected the 6-Frame video feature (64×64×3) with the 181-dimensional audio spectrum (MFCC/Me spectrogram) and input it to CNN+ConvLSTM, but faced the problem of dimensional disaster [17]. Jiang, Z. et al. verified the limitations of early multimodal fusion. The direct mosaic of visual (VGG19 emotional classification), acoustic (PyAudioAnalysis), and linguistic (Roberta) features led to an AUROC of only 0.59 T. The Q-Former module in the article of Li, Y. et al. learns visual and language attention mapping attention fusion, which enhances the accuracy of multimodal fusion results through dynamic weighting [12]. The DJVAE framework in Meng, T. et al. achieves 69.27% WAF1 in the IEMOCAP dataset by jointly optimizing the reconstruction loss and KL divergence through the variational autoencoder [15]. Middleya, A. I. et al. further proved that the model level fusion (splicing flat features+full connection layer) is 5-10% higher than the decision level voting accuracy [17]. Decoupling learning embodies a new trend. The FDMER framework of Yang et al. disentangles features into a modality-invariant subspace (aligned via adversarial learning with angular margin loss) and modality-specific subspaces (enforcing independence via HSIC constraint). These representations are fused by a Cross-Modal Attention Fusion (CMAF) module that learns adaptive weights through cross-modal interaction and attention weighting, achieving 84.6% Acc2 on CMU-MOSEI (BERT-based model) [18]. Li, B. et al. proposed a Dual-level Disentanglement Mechanism (DDM), Modality-level Contrastive Learning to enhance the clustering of similar features, Utterance-level Alignment of multimodal semantics, and TCP-based Dynamic Weighting in combination with Contribution-aware Fusion Mechanism (CFM) (such as focusing on voice features when emotions fluctuate) [19]. For physiological signal fusion, Qin, S. M. adopts an integral-based judgment mechanism, setting 9 independent thresholds such as P300 amplitude ≥8μV and ER-SCR peak ≥ 0.3, scoring the standard features, and a total score ≥4.5 to determine lying [6]. Ernst, H. et al. used logistic regression to screen 8 core biomarkers (QTVi/LVVI/PTA) and constructed an acute stress detection model with a sensitivity of 78.0% and specificity of 97.6% [8]. The COGMEN model in Joshi et al.'s article uses graph neural networks (specifically Relational GCN followed by Graph Transformer) to model speaker relationships, with intra-speaker edges capturing dependencies within the same speaker's utterances and inter-speaker edges capturing stimulus-response patterns across different speakers [20].

3. Typical Applications and Performance Comparison

3.1. Overview of Typical Applications and Performance Comparisons of Multimodal Perception

Multimodal perception technology has shown wide application potential in the field of mental state detection and analysis. Its core is to integrate multiple modal information, such as physiological signals, behavior data, voice, vision, and text, so as to make up for the limitations of single mode and improve the comprehensiveness and accuracy of detection. In the field of health care, multimodal systems are widely used in mental health assessment. For example, the remote video interview described by Jiang, Z. et al. in the article combines face, voice, language and cardiovascular signals (RPPG) to effectively screen depression (MDD) and anxiety (GAD-7) through the late fusion strategy. The AUROC is up to 0.82, which is significantly better than the single-mode method [7]. Cohen, J. et al. further implemented decision fusion for suicide risk identification using percent pause time (PPT) and facial movement features, elevating the AUC to 0.76. This underscores multimodality's advantage in detecting subtle behavioral alterations, particularly manifestations like psychomotor retardation observed in depression [11]. In the security and monitoring scenario, the lie detection system developed by Qin, S. M. integrates EEG, GSR, BCG and temp signals. Based on the concealed information test (CIT) paradigm, it achieves an accuracy rate of 85% through the 9-item feature

integration model, and the characteristic change rate of the criminal group is more than 80%, verifying the robustness of physiological signal complementarity (such as P300 amplitude rise and SDNN decline) in lie detection [6]. Similarly, Guo, X. et al.'s deception detection framework uses audiovisual mode, combined with UT adapter and PAVF module, to achieve an accuracy rate of 66.84% on the DOLOS data set, especially through facial action units (such as chin lifting) and voice spectrum changes to enhance its practicability in lie detection [10].

In emotion recognition and daily behavior analysis, multimodal technology optimizes human-computer interaction and education scenes through a dynamic fusion mechanism. Li, B. et al. and Joshi, A. et al. focused on multimodal emotion recognition in conversations, utilizing text, audio, and video modalities. By employing contextualized graph neural networks (GNN) and modality disentanglement strategies, they achieved an F1-score of 67.63% on the IEMOCAP dataset, representing an improvement of nearly 5% over unimodal approaches [19] [20]. The model-level fusion framework for audio-visual modalities proposed by Middya et al. achieves 86% accuracy on the RAVDESS dataset. By leveraging ConvLSTM2D to capture spatio-temporal dynamics in facial expressions (e.g., pupil constriction), this approach effectively overcomes the limitations of static models that ignore temporal variations [17]. In the context of child health management, Jiang, Z. et al.'s Health Prism system integrates contextual data (e.g., sleep patterns, family income) and motion sensor data (tri-axial accelerometer) to infer six-dimensional health indicators through a gate mechanism (HPM model), achieving a mean AUC of 0.807. The influence of motion data on physical activity intensity (MVPA) and connectedness (CONN) is visualized via heatmaps and temporal analysis, providing data-driven support for personalized interventions [16]. Additionally, the rumor detection method (MRML) proposed by Peng, L. et al. and the fake image recognition approach using Large Multimodal Models (LMMs) developed by Li, Y. et al. extend the application of multimodal analysis in social media analytics. The former achieves an accuracy of 89.7% on the Weibo dataset by leveraging text-image cross-modal consistency verification, thereby enhancing the validity of misinformation identification [9] [12].

In the diagnosis of mental illness and stress assessment, multimodal perception reveals pathological mechanisms through high-dimensional data fusion. The acute stress framework in the article by Ernst, H. et al. integrates psychometric measures (SAM scale), chemical biomarkers (cortisol), and bio signals (ECG, EDA), identifying 8 core parameters such as QT variability and electrodermal activity, with a sensitivity of 78.0% and specificity of 97.6%, covering multi-pathway responses including sympathetic activation and respiratory drive [8]. The MRI multimodal fusion (fMRI and sMRI) in Liu, R. et al.'s article achieved an accuracy of 75.1% in the diagnosis of autism spectrum disorder (ASD) through the MGF mechanism. Key biomarkers such as reduced volume of the left superior temporal gyrus, through decoupled visualization, provide an interpretable basis for precision medicine [14]. In general, the common point of these application scenarios is to use multimodal complementarity, such as physiological signals (EEG, HRV) to provide objective biological indicators, visual and audio modes to capture behavior dynamics, and text and language data to analyze cognitive patterns, so as to achieve end-to-end optimization in complex mental state analysis.

3.2. Summary of Multimodal Perception Performance Comparison

Table 1 summarizes the key performance indicators reported in various articles, covering the main application fields, model methods, modal combinations and performance data. The table design is based on actual data to ensure specificity and comparability. Performance indicators include accuracy (ACC), AUC, F1 score, etc., to show the comparative advantages of multimodal fusion.

Table 1. Performance comparison of multimodal perception [13]

Application area	Model/method	Modal combination	Key performance indicators
Lie detection	Multimodal integration model	EEG, GSR, BCG, Temp	Accuracy rate>85%, change rate of crime group characteristics>80%
Mental health assessment	Late fusion strategy	Face, voice, language, cardiovascular (RPPG)	Auroc: 0.82 (depression detection), AUC: 0.76 for suicide risk identification after fusion
Acute psychological stress assessment	Self-Optimizing Logistic Regression	Psychometric, cortisol, ECG, EDA	Sensitivity: 78.0%, specificity: 97.6%, F1 score: 0.82
Rumor detection	Metric Learning (MRML)	Text, image	Accuracy: 89.7% (Weibo), F1: 0.892
Deception detection	Cross-modal Learning (PECL)	Audio, visual	Accuracy: 66.84%, ACC improved to 66.84% after multi task fusion optimization
Forged image detection	Large Multimodal Models (LMMs)	Visual, text	False image detection accuracy: 59.87%, real image: 96.20%
Emotion recognition (Laboratory)	Feature level fusion	EEG, GSR, ECG	Accuracy: 93.6% (DEAP dataset), AUC: 0.98
Diagnosis of mental illness	Multi-head Gated Fusion (MGF)	fMRI, sMRI	Accuracy: 75.1% (ASD), 87.2% (SCZ), AUC: 0.91
Dialogue emotion recognition	CBERL (Class Boundary Enhanced Representation Learning)	Text, audio, video	WAF1: 69.2% (IEMOCAP), Minority F1 increased by 10-20%
Analysis of children's health	Hybrid Prediction Model with Gating Mechanism (HPM)	Context data, motion sensor	Average AUC: 0.807, MVPA recognition AUC: 0.781
Audio visual emotion recognition	Model level fusion	Audio, visual	Accuracy: 86% (ravless), f1: 0.99 (savee)
Emotion Recognition (with Disentangled Learning)	FDMER	Text, audio, visual	MAE: 0.724 (CMU-MOSI), Corr: 0.773
Dialogue emotion recognition (context)	DF-ERC	Text, audio, video	W-F1: 67.03% (MELD), 0.32% higher than baseline
Emotion Recognition with Graph Neural Networks (GNN)	COGMEN	Audio, text, video	F1-score: 67.63% (IEMOCAP), 4-class Acc: 84.50%

3.3. Key Performance Discovery and Multimodal Advantages

The performance improvement of multimodal perception is mainly due to modal complementarity and fusion strategy innovation. In terms of modal complementarity, physiological signals (such as the EEG P300 amplitude of Qin, S. M., and the QT variability of Ernst, H. et al.) provide highly specific biomarkers, but are vulnerable to environmental noise [6] [8]. Behavioral modality (such as the facial action unit studied by Guo, X. et al., and the speech spectrum in the article by Middleya, A. I. et al.) enhances real-time performance, but relies on high-quality acquisition [10] [17]. Fusion strategy becomes the key. Early fusion (feature stitching) is effective in simple tasks (Middleya, A. I. et al.'s audio visual ACC 86%), but high-dimensional data is prone to dimensional disasters (Jiang, Z. et al.'s article on early fusion AUC is reduced to 0.63) [16] [17]. Late fusion (decision voting or weighting) has significant advantages in complex scenes. For example, the AUC of mental health

assessment of Jiang, Z. et al. has been increased to 0.82, and the AUC of suicide risk has been increased from 0.65 to 0.76 by Cohen, J. et al [11]. Innovative fusion mechanisms like Liu, R. et al.'s MGF (Multiheaded Gating Fusion) and Yang, D. et al.'s FDMER (Feature-Disentangled Multimodal Emotion Recognition) optimize performance through dynamic weight allocation (e.g., modality contribution learning), achieving an ACC of 87.2% in mental disorder diagnosis – a 2-3% improvement over conventional methods [14] [18].

Performance comparisons reveal core trends. Multimodal systems are significantly better than Single-Mode Systems in scenarios with sufficient data. For instance, Jiang et al. reported an increase in the AUC for children's health analysis from 0.784 (using a single-mode model) to 0.807 (using their multimodal model) [16]. Similarly, Joshi et al. achieved an improvement in the F1-score for dialogue emotion recognition, from 63.4% (text-only) to 67.63% (multimodal) [20]. However, performance is significantly influenced by dataset characteristics. In balanced datasets, such as the Weibo rumor detection study by Peng, L. et al., multimodal ACC reached 89.7%. In contrast, for imbalanced data like the MELD dataset in Meng, T. et al. (where the "fear" class constitutes only 1.91%), integrating data augmentation (GAN-based generation) improved the minority-class F1 to 22.2% [9] [15]. In terms of computational efficiency, lightweight models, such as the ConvLSTM2D with only 3 million parameters in the work by Middya et al., support real-time applications, while complex frameworks, like the RGCN in the study by Li et al., improve accuracy (W-F1 67.03%) but face challenges in online inference [17] [19]. Modality importance analysis reveals that text dominates cognitive-related tasks, where removing text increased MAE to 1.275 [18], while visual and audio modalities exhibit heightened sensitivity to behavioral dynamics. In Guo, X. et al., the weight of the video modality increased by 19% in their framework [10].

4. Limitations and Future Directions

Although the detection and analysis of mental state based on multimodal information have made significant progress in theory and application, their limitations still hinder large-scale deployment. The current research is generally faced with the problems of small sample size and insufficient diversity, resulting in over-fitting and poor generalization of the model. At the same time, hardware systems such as EEG acquisition equipment are highly dependent on a low-noise static environment and have weak environmental adaptability. Although the anti-interference design optimizes the RF protection through the TVs tube and RC network, it cannot completely eliminate the interference of dynamic scenes. At the data level, modal asynchrony, such as audio-visual signal delay, and emotional category imbalance, such as the MELD dataset, "fear" category only accounts for 1.91%, which aggravates the recognition error, and the F1 score of a few categories is often lower than 11.2%. In addition, the computational efficiency has become a bottleneck. The deep fusion model, such as the COGMEN architecture, has a large number of parameters and high real-time reasoning delay, which is difficult to meet the clinical or security needs. Ethical privacy risks can not be ignored. Although remote interview systems, such as the Tina platform, improve accessibility, the Zoom platform data collection may leak sensitive information, and the black box decision-making mechanism lacks transparency, which can easily trigger diagnostic bias. These limitations jointly limit the practicability and reliability of the technology.

The future direction needs to focus on technological innovation and system optimization to realize the transformation from the laboratory to the real world. Expanding the sample size and enhancing generalization are the primary tasks. It is recommended to cover multi-age and ethnic groups through multi-center clinical validation and transfer learning. Deepen the research of the system interaction mechanism combined with dynamic network physiology. Environmental adaptability and real-time deployment can be optimized by lightweight hardware and an adaptive algorithm. Edge computing deployment can improve the robustness of dynamic scenes. Multimodal fusion strategy should give priority to late fusion to avoid feature redundancy, and extend to physiological signals (such as EEG and RPPG) and time series modeling (such as HMM to capture heart rate dynamics). Self-supervised

learning can reduce label dependency and improve model efficiency. Cross-domain applications such as mental health monitoring (dynamic tracking of depression) and intelligent driving safety need to integrate an AI ethical framework (such as privacy protection design). Tina, a cloud system, has demonstrated the potential of telemedicine. Standardization is the ultimate goal. A unified evaluation framework (such as the combination of "QT variability+respiration+dermatoelectricity") will promote clinical integration through multi-center verification to ensure reliable and scalable technology.

5. Conclusion

This study provides a systematic review of research progress in detecting and analyzing human mental states using multimodal information. By integrating non-contact sensing, contact-based physiological signal acquisition, and cross-modal fusion techniques, this study has constructed a comprehensive output framework spanning from information data collection to decision fusion.This work systematically investigates and implements a multimodal mental state detection and analysis framework, covering the entire process, including data acquisition, feature fusion, typical applications, and challenges and future directions. Multimodal approaches significantly enhance detection accuracy and robustness through deep integration of physiological and behavioral signals, while also delivering breakthroughs in interpretability.

With the continuous research and development of multimodal information processing and intelligent perception technologies, mental state detection in humans will demonstrate greater application potential in cross-scenario transfer and personalized intervention.Future research must strike a balance between data privacy protection and cross-domain sharing to advance the feasibility of methods in real-world complex scenarios. Technological breakthroughs in sensor comfort and wearability are also essential to minimize data inaccuracies caused by device discomfort. Simultaneously, the introduction of causal reasoning and methods to enhance interpretability will become a key direction for improving model transparency and trustworthiness, providing a reliable theoretical foundation for their standardized application in high-risk fields such as healthcare and the judicial system.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] World Health Organization. World mental health report: Transforming mental health for all. Geneva: World Health Organization, 2022.

[2] Global Wellness Institute. Global Wellness Economy Monitor 2023. Miami, FL: Global Wellness Institute, 2023.

[3] Fang, X., Liu, W., & Kawakami, K. Evaluating past emotions in changing facial expressions: The role of current emotions and culture. Emotion, 2024, 24 (1): 213 – 224.

[4] Ma, J., Chen, X., Huang, J., et al. Cam4DOcc: Benchmark for camera-only 4D occupancy forecasting in autonomous driving applications. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024: 21486 - 21495.

[5] IEEE. IEEE conference publication (TAFFC.2024.3387654). In IEEE Conference Proceedings. IEEE, 2024.

[6] Qin, S. M. Development of multimodal physiological signal acquisition system and its application in polygraph. Master's thesis, Lanzhou University, 2022.

[7] Jiang, Z., Seyedi, S., Griner, E., Abbasi, A., Rad, A. B., Kwon, H., ... & Clifford, G. D. Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns. IEEE journal of biomedical and health informatics, 2024, 28 (3): 1680 - 1691.

[8] Ernst, H., Scherpf, M., Pannasch, S., Helmert, J. R., Malberg, H., & Schmidt, M. Assessment of the human response to acute mental stress–An overview and a multimodal study. PLoS One, 2023, 18 (11): e0294069.

[9] Peng, L., Jian, S., Li, D., & Shen, S. Mrml: Multimodal rumor detection by deep metric learning. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1 - 5.

[10] Guo, X., Selvaraj, N. M., Yu, Z., Kong, A. W. K., Shen, B., & Kot, A. Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 22135 - 22145.

[11] Cohen, J., Richter, V., Neumann, M., Black, D., Haq, A., Wright-Berryman, J., & Ramanarayanan, V. A multimodal dialog approach to mental state characterization in clinically depressed, anxious, and suicidal populations. Frontiers in psychology, 2023, 14: 1135469.

[12] Li, Y., Liu, X., Wang, X., Lee, B. S., Wang, S., Rocha, A., & Lin, W. Fakebench: Probing explainable fake image detection via large multimodal models. IEEE Transactions on Information Forensics and Security, 2025.

[13] Liu, H., Lou, T., Zhang, Y., Wu, Y., Xiao, Y., Jensen, C. S., & Zhang, D. EEG-based multimodal emotion recognition: A machine learning perspective. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1 - 29.

[14] Liu, R., Huang, Z. A., Hu, Y., Zhu, Z., Wong, K. C., & Tan, K. C. Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35 (6): 7627 - 7641.

[15] Meng, T., Shou, Y., Ai, W., Yin, N., & Li, K. Deep imbalanced learning for multimodal emotion recognition in conversations. IEEE Transactions on Artificial Intelligence, 2024.

[16] Jiang, Z., Chen, H., Zhou, R., Deng, J., Zhang, X., Zhao, R & Ngai, E. C. Healthprism: a visual analytics system for exploring children's physical and mental health profiles with multimodal data. IEEE Transactions on Visualization and Computer Graphics, 2023, 30 (1): 1205 - 1215.

[17] Middya, A. I., Nag, B., & Roy, S. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Knowledge-based systems, 2022, 244: 108580.

[18] Yang, D., Huang, S., Kuang, H., Du, Y., & Zhang, L. Disentangled representation learning for multimodal emotion recognition. In Proceedings of the 30th ACM international conference on multimedia, 2022: 1642 - 1651.

[19] Li, B., Fei, H., Liao, L., Zhao, Y., Teng, C., Chua, T. S., ... & Li, F. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In Proceedings of the 31st ACM International Conference on Multimedia, 2023: 5923 - 5934.

[20] Joshi, A., Bhat, A., Jain, A., Singh, A., & Modi, A. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 4148 - 4164.