Analysis of Mental Health Assessment Methods Based on Multimodal Signals

Jie Shao *

Faculty of innovation Engineering, Macao University of Science and Technology, Macao, China

* Corresponding Author Email: 1240029367@student.must.edu.mo

Abstract. According to the World Health Organization, there are more than 300 million people with depression in the world. And if this trend is not controlled, it is expected that depression will become the most serious and common disease in the world in 2030, and the mental health problems of college students are particularly serious. In order to effectively detect mental health, this review analyzes and summarizes each of the three commonly used multimodal architectures at this stage, compares their typical studies and experiments to explore the demand relationship between multimodal signals and architectures, explores the advantages and disadvantages of architecture algorithms, gives new algorithm optimization schemes, and gives reasonable suggestions for new application environments. The results show that the three multimodal architectures explored have higher accuracy than the single-modality. However, there are deficiencies in anti-interference, dataset collection, and scenario interaction, which can be optimized by dataset update and lotus effect algorithm.

Keywords: Multimodal Architecture; Multimodal Signaling; Mental Health; Emotion Recognition.

1. Introduction

With the growth of the population and the advancement of the medical field, the diagnosis and treatment of physical diseases have made a qualitative leap. However, the neglect of mental health has become one of the most serious hidden problems in the global situation. According to the analysis of the "China National Mental Health Development Report (2021~2022)", the detection rate of depression risk among college students reaches 21.48%, and the detection rate of anxiety risk can be as high as 45.28%. In addition, young people aged 18 to 35 account for far more than 40% of patients with mental disorders. In order to effectively protect human mental health, monitoring human emotions through modal indicators has become the focus of research and an important means. Traditional unimodal research only uses a fixed signal (such as a single visual signal, single electroencephalogram (EEG) signal) to monitor the accuracy and versatility. For example, if a depressed patient loves to smirk, it is difficult to judge the difference between him and a healthy person. However, the multimodal fusion architecture has greatly improved its monitoring accuracy and versatility compared with single-modal due to the weight proportion of multiple information. Therefore, multimodal architecture has become the focus of current research.

There have also been key breakthroughs in research on depression, an extremely serious mental health problem. In the past, it has been criticized for its costly and difficult to detect. Ricardo Flores et al. also proposed a new multimodal deep learning architecture, WavFace, which results in a higher accuracy by inputting audio and facial features and aligning the use order and spatial self-attention [1]. The experimental research of Dong et al. proposed a new computing network architecture based on three architectures, namely a hierarchical attention network based on memristors, which innovated a low-cost, low-power, and wearable mental health detection system in the field of multimodal affective computing [2]. Yuichi Ishikawa1 et al. proposed a method based on physiological signal emotion recognition, which proposed to separate cross-modal sentiment factors and modal-specific interference factors on the original infrastructure, which significantly improved the accuracy of emotion recognition and was conducive to better detection of mental health [3]. Jiang et al. also proposed a new interactive model, HealthPrism, which has a gate mechanism and has higher performance in processing cross-modal data [4].

By sorting out and analyzing the previous multimodal research, this study divides multimodal architectures into three categories according to their characteristics and the demand for different types of multimodal information: physiological signal-oriented framework, real-time intervention-oriented framework, and non-contact universalization framework. These three frameworks lead the application of inquiry in three different directions, and also divide modal information into three types. This study will extend it to make the three frameworks better applicable to practical scenarios.

2. Comparative Analysis of Different Model Signals and Their Frameworks

2.1. Application of the Physiological Signal-Guided Framework

The physiological signal-oriented framework mainly focuses on the accurate identification of mental disorders as the core goal, and focuses on the application and exploration of multimodal biological signals. The experiments of Imran Mehmood et al. well demonstrate the accuracy of physiological signals in this framework [5]. The research goal of the experiment is to evaluate the mental health of workers through a multimodal fusion model. The dataset used was 16 male excavator operators, all of whom had experience in construction sites. The experimental environment is set up, simulating the real construction site scene and the excavator is equipped with sensors, which are deployed in Table 1 below.

Table 1. Sensor deployment

sensor	Mounting location	Sampling parameters
EEG Headband (Muse)	Operator Head (AF7/AF8/TP9/TP10)	4 channels, 256HZ
Camera	Internal test of the front windshield of the cockpit	1440×1440 pux, 30 fuchs
EDA Watch (Empatica E4)	Operator's wrist	Single channel, 4HZ

Table 2. Accuracy of the results of each modal combination

Modal data sum	DT accuracy	Optimal feature contribution
EEG+EDA+FF	**96.2%**	Facial features (FF) are the most sensitive to fatigue states
FF+EDA	**96.9%**	EDA was significantly differentiated in the late stage of fatigue (physiological stress).
EEG+FF	**97.1%**	The ratio of theta to alpha waves in EEG is associated with cognitive load.
EEG + EDA	**85.0%**	A single sensor has the lowest performance and can be used to validate the need for multimodal detection

The experimental process is as follows. Set a repetitive and non-work-demanding task for workers, that is, dig the soil and load it into a transport vehicle, and so on for 60 minutes, and the design is based on the fact that time is the most effective way to induce fatigue and produce bad emotions for workers. Data acquisition is synchronized through real-time transmission of EEG signals via Bluetooth to record the ratio of power in different bands. The EDA signal is used to test the conductivity reaction of the skin, extracting five features such as mean and standard deviation, and the visual screen signal is segmented frame by frame by OpenCV to observe the geometric features. The standard of fatigue annotation score is 20 minutes, 40 minutes, and 60 minutes at the beginning of the time mission. Finally, a physiological signal-oriented framework is constructed by machine learning model, and the calculation accuracy is shown in Table 2.

Compared with the results of single-modal EEG assessment, the accuracy of its use alone is only 73.8%. The experimental conclusion is that the architecture accuracy of DT plus three modalities is the highest, and the experimental effect is the best. The above experiments are typical structures and experiments of physiological signal-oriented frameworks, and the core of which is the monitoring of

physiological information. This study focuses on the acquisition and combination of physiological modal signals, generates a multimodal dataset, and then judges the psychological state by selecting different machine learning models. The monitoring accuracy after the final modal combination is much higher than that of the single modality, and the accurate capture and detection of physiological signals has an important application prospect in the application of mental health in construction sites and clinical medicine.

2.2. Real-Time Intervention-Oriented Framework

The core focus of the real-time intervention-oriented framework is on dynamic monitoring and feedback in everyday situations. It's mainly about collecting various types of information in real time. This framework was well demonstrated in the experiments conducted by TENG GUO and others[6]. The experimental process follows data collection, multimodal representation features, data augmentation, and training of the mental health detection model. The dataset originates from 509 freshmen students at Chinese universities, and after excluding invalid data, a remaining sample of 584 students is used. The labels for mental health are assessed using the Symptoms Checklist-90 (SCL-90) self-assessment scale. Academic performance is recorded from the Learning Management System (LMS) for all courses in the first semester. Demographic information such as age and ethnicity is obtained from the LMS. Appearance modality data is collected through student photographs taken with the Fujifilm FinePix S5 Pro camera. Social networks are gathered through surveys administered to students. The core principle of multimodal feature representation algorithms is to convert heterogeneous data into low-dimensional vectors. Social life uses the MOON algorithm, that is, a multi-view social network .

$$G = (U, V\{E(v): v \in V\}) \tag{1}$$

U is the set of student nodes, V is the 8 views of the social scene (obtained through the questionnaire), and E(v) is the edge set of view v. Embedded strategy is adopted in three types of node pairs. As shown in Table 3.

Table 3. Correspondence of View Node Types and Their Representations

Node Type	Mathematical Representation	Physical significance
View Node Pair	$(i^{(v)}, j^{(v)}) \in \Omega^{(v)}$	Preserve the single-view topological structure.
Cross-view same-node pairs	$(i^{(v_0)}, i^{(v')})$	Aligning the embeddings of the same node across different views (first-order relationship)
Cross-view cross-node pairs	$(i^{(v_0)}, j^{(v')})$	Guiding the target view embedding (second-order relations) using the source view.

Optimizing the complexity of algorithms through loss functions. The representation of appearance features is implemented using the CNN autoencoder algorithm.

$$L_{AE} = \|x - Dec(Enc(x))\|^2 \tag{2}$$

The encoder Enc() is used to extract latent features of appearance, while the decoder Dec() is used to reconstruct the image.

The academic performance is addressed using a variant of the One-Hot Autoencoder method. To tackle the issue of label imbalance, data augmentation through the SMOTE algorithm has been additionally incorporated. The principle involves performing interpolation of minority class samples within the feature space. The model verification and training primarily utilize a three-layer DNN classifier, consisting of an input layer, a hidden layer, and an output layer. The input layer represents a 23-dimensional feature vector, encompassing aspects such as social life, appearance, academic performance, and demographics. A key design consideration is the unification of heterogeneous features to avoid numerical discrepancies. The hidden layer employs techniques to prevent overfitting, with the core mechanism being Dropout (with a ratio of 0.3 being optimal), which randomly drops

30% of the neuron inputs during data training. The principle behind this is to compel the network not to rely on specific neurons, thus enhancing generalization. It is analogous to indirectly training multiple subnetwork ensembles, normalizing the process, and accelerating convergence to improve learning capability.

A new real-time intervention-oriented framework called MindMatrcis was also proposed in the study by Swapnil Joshi et al [7]. Its core is similar to that of the previous study, as it achieves emotion recognition and mental health assessment by integrating questionnaire, video, and text data. The modal data module focused on here primarily involves four methods of data acquisition. The questionnaire was obtained through the DASS-21 scale, audio emotion recognition was acquired using CNN-LSTM with Librosa, video emotion information was obtained through 2D CNN with OpenCV, and emotional diaries were collected using BERT for text sentiment analysis. Its core function primarily involves quantifying levels of stress and depression for scoring purposes, as well as analyzing various emotions through voice recognition. The enhancement of the model's generalization ability is achieved through data augmentation based on micro-expression changes. Finally, the system generates an emotional trend report, identifying long-term psychological patterns and implementing dynamic long-term real-time monitoring.

By balancing the modal weights of different modalities in daily life through the above two studies, it is possible to achieve real-time monitoring of the emotional and mental states reflected behind the daily behaviors of students or corporate employees, and to provide corresponding feedback in a timely manner, making interactive therapy more prompt. This long-term tracking mechanism has shifted mental health from passive treatment to proactive prevention, showcasing a wide range of application prospects in social-level mental health monitoring.

2.3. Non-contact Universalization Framework

The non-contact universalization framework primarily targets low-burden and highly accessible scenarios, utilizing ordinary cameras and microphones to achieve seamless interaction. According to Rajesh Titung's research, machine interaction learning plays an important role in solving the lightweight modal perception and sentiment analysis of data scarcity [8]. The research process is as follows: the validation experiment of emotion-inducing stimuli, with the primary goal of collecting reliable multimodal data for the subsequent iML framework, and validating whether specific stimuli can effectively induce the target emotions (initially focusing on frustration and surprise). Among them, the subjects were human participants, and the elicitation tasks utilized carefully designed stimuli to attempt to provoke the target emotions. The data collection occurred synchronously while the participants executed the elicitation tasks, encompassing multimodal data including language (viewed as a baseline condition), as well as language + x such as eye tracking, speech prosody, and facial expressions. Through self-reporting for self-assessment and partner reports (where participants evaluate each other), the differences between the two assessments are validated to achieve labeling. The data collection and preliminary modeling within the IML framework aim to develop an interactive interface that allows interaction with stimuli, ultimately resulting in the generation of multimodal data streams. Such iterative cycles, carried out multiple times, will utilize the seed dataset obtained from testing to be used in conjunction with the foundational learning of the training framework. Expand the dataset of lightweight design based on language text by transforming lightweight modal information into strongly correlated modal information during the process of interactive learning.

In addition to verifying the superiority of the lightweight design of the universal framework through human-computer interaction, there are also some large model architectures that can directly demonstrate and evaluate the uniqueness of this architecture, such as a multimodal deep learning framework called FusionNet proposed by Shifa Fathima I et al. [9]. Its core purpose is to integrate various lightweight information sources such as text, speech, and visual data in order to comprehensively understand and recognize emotions and detect mental health. The core innovation of this research lies in the integration of three lightweight modalities: text, language, and vision, for

the first time. The deep learning architecture employs CNN (Convolutional Neural Networks) to process variations in facial expressions, thereby capturing the visual modality, LSTM/RNN (Recurrent Neural Networks) to process audio signals for the acquisition of the speech modality, and finally utilizes Transformers to analyze semantics and emotions. Fusion strategies employ three approaches: feature-level fusion involves directly concatenating the low-level features of different modalities as input to the classifier; decision-level fusion centers on independent predictions from each modality, which are later combined; and hybrid fusion integrates the strengths of the former two by dynamically weighting important features. The main focus is on the attention mechanism, which dynamically allocates weights to concentrate on key information. The datasets used in the experiments include IEMOCAP, MELD, RAVDESS, and AffectNet among other facial images, and the final experimental results, along with a comparison of the performance of each fusion method, are presented in Table 4.

Table 4. Comparison of Performance of Different Fusion Models

Model Type	Accuracy Rate	F1 value and score
Monovision	**85.0%**	**83.0%**
Single Language	**80.0%**	**78.0%**
Single Text	**78.0%**	**76.5%**
Feature Fusion	**92.5%**	**91.0%**
Decision Integration	**91.0%**	**90.0%**
Mix and Converge	**93.5%**	**92.5%**

This mechanism of dynamic decision integration significantly enhances the accuracy of emotion identification and provides insightful implications for the lightweight design of simple modalities.

Both studies mentioned above have validated the significant value of lightweight modalities in mental health assessment. By employing simple modalities to accurately identify emotions to the greatest extent possible, and integrating dynamic fusion with machine learning, the precision of mental health detection can be enhanced. This approach also allows for a more universal and comprehensive evaluation of mental health.

3. Limitations and Prospects for Application

3.1. Limitations and Prospects of the Application of Physiological Signal-oriented Architecture

3.1.1 Limitations of the Application of Physiological Signal-oriented Architecture

The small scale of the dataset and its insufficient versatility are specifically reflected in the fact that the dataset only includes workers from a single type, without extending to workers across various production lines. To further the research, expanding the dataset from only 16 excavator operators to include laborers from half of the production line on a typical construction site or a broader variety of workers would enhance the comprehensiveness of the data. The lack of algorithmic robustness is specifically reflected in more precise clinical diagnoses, where the algorithms exhibit weak noise resistance. In clinical diagnoses or on construction sites, the stability is still insufficient, which can be improved through the optimization of gating mechanisms or the application of the lotus effect algorithm. Elham Dalirinia et al. proposed a very constructive noise reduction method, namely the multimodal lotus effect algorithm M-LEA [10]. Current research on this algorithm is primarily applied to robotic obstacle avoidance tasks in high-noise factory environments. Its core is to evaluate stability through the independent evolution of subgroups and the population growth rate (GM), which aligns closely with the self-attention weighting mechanism commonly used in multimodal mental health diagnostics. Both methods adjust resource allocation based on real-time evaluation results, and research indicates that the multimodal lotus effect algorithm demonstrates significantly better noise

resistance than sub-attention weighting while requiring lower computational power. In the future, this algorithm can be utilized to address the substantial impact of noise on this experiment.

3.1.2 Prospects for Application of the Application of Physiological Signal-oriented Architecture

This framework has a comprehensive mental health monitoring system, which can be applied in some practical scenarios, such as stress-related mental health assessments for high-risk professions, including doctors, firefighters, and long-distance bus drivers. Adaptive interventions can also serve as a crucial aspect of their application. When the system detects a sustained state of stress, it automatically triggers relaxation prompts (such as guided breathing) or adjusts task allocation, aiming to regulate workers' long-term mental health to a stable state and enhance their work efficiency. In the monitoring of clinical depression, it can also play a significant role, such as in the early warning of suicidal tendencies through auxiliary indicators of depression, integrating multimodal data to identify high-risk behavioral patterns, low activation of the prefrontal cortex in EEG, absence of EDA responses, and facial expressions. Precise prevention and treatment, combined with GPS positioning, and linkage with community psychological service institutions for active intervention.

3.2. Limitations and Application Prospects of Real-time Intervention-oriented Frameworks

3.2.1 Limitations of Real-time Intervention-oriented Frameworks

The main issue with both studies is that the accuracy of the datasets is clearly insufficient and lacks interpretability. The origin of the datasets directly stems from survey questionnaires, which do not guarantee the accuracy of the data. Secondly, the modal selection is somewhat one-sided, as the choice of a multimodal architecture that integrates social networks, academic performance, and physical appearance is not sufficiently persuasive. It could also include behaviors related to internet usage and patterns of online activity, which are often overlooked as implicit modal information.

3.2.2 Prospects for Application of Real-time Intervention-oriented Frameworks

This architecture can be applied in more life-oriented scenarios such as schools and enterprises, focusing more on real-time monitoring in daily life, as opposed to the application of physiology signal-oriented architecture, which emphasizes timely feedback. For example, the creation of a personalized mental health intervention system involves a comprehensive process of dynamic risk alerting, real-time analysis of students' abnormal behaviors, and emotional anxiety. This system then automatically triggers a graded warning, recommending the involvement of counselors and professional psychological consultants. Finally, timely and precise intervention plans are generated, categorizing students into different types such as socially isolated, academically pressured, and foreign trade anxious. The aim is to ensure psychological well-being in people's lives.

3.3. Limitations and Application Prospects of Non-contact Universality Framework

3.3.1 Limitations of Non-contact Universality Framework

Human-computer interaction is a long-term learning process involving labeling, and therefore lacks real-time attributes. The output results from this learning represent a lengthy research process, characterized by latency. The research summarized in this review on the first non-contact universality framework is merely a conceptual study, lacking a rigorous proof process. The data collection inputs and outputs adopted by the two research institutes lack comprehensive privacy protection measures, resulting in insufficient versatility in certain confidential settings.

3.3.2 Prospects for Application of Non-contact Universality Framework

Due to the currently limited resources in mental health care, many patients are unable to attend in-person consultations and treatments in time, thus missing the optimal period for mental health diagnosis and treatment. The lightweight architecture and data augmentation mechanisms have significantly enhanced the quality of remote medical services, which initially had a low accuracy rate in diagnosis. The lightweight design of this architecture can also be integrated into the first two

architectures, allowing it to support mobile and embedded devices for long-term monitoring in future applications, thereby assisting in the implementation of intervention-oriented frameworks.

4. Conclusion

The review summarizes the core experiments and important breakthroughs of three significant architectures in the field of multimodality: the physiological signal-oriented framework, the real-time intervention-oriented framework, and the non-contact universal framework. It emphasizes the preferences and demands of these three architectures for different types of multimodal information and analyzes their limitations and potential applications in detail. Moreover, the relationships among these three frameworks are not contradictory; instead, they are complementary and intersecting. The three types of frameworks exhibit significant functional complementarity—physiological frameworks provide pathological-grade biological markers, real-time frameworks drive immediate feedback, and non-contact frameworks achieve seamless and universal coverage. The overview focuses more on the complete experimental processes and practical application scenarios of the three architectures, and does not delve into detailed comparisons of different algorithms or algorithm optimization. Future research directions may further concentrate on the iterative processes of the three algorithms and algorithm optimization, while also introducing more novel modal information, such as behavioral modal and other abstract modes that are not commonly addressed.

References

[1] Flores R, Tlachac M, Shrestha A and Rundensteiner E A, WavFace: A Multimodal Transformer-Based Model for Depression Screening. in IEEE Journal of Biomedical and Health Informatics, 2025, vol. 29, no. 5, pp. 3632-3641.

[2] Dong Z, Ji X, Lai C S, Qi D, Zhou G and Lai L L, Memristor-Based Hierarchical Attention Network for Multimodal Affective Computing in Mental Health Monitoring, in IEEE Consumer Electronics Magazine, 2023, vol. 12, no. 4, pp. 94-106

[3] Ishikawa, Y et al. Learning Cross-Modal Factors from Multimodal Physiological Signals for Emotion Recognition. PRICAI : Trends in Artificial Intelligence. 2023, vol. 14325. Singapore: Springer, 2023. 438–450.

[4] Jiang Z et al., HealthPrism: A Visual Analytics System for Exploring Children's Physical and Mental Health Profiles with Multimodal Data, in IEEE Transactions on Visualization and Computer Graphics, 2024, vol. 30, no. 1, pp. 1205-1215.

[5] Mehmood, I, Li, H, Umer, W, Arsalan, A, Anwer, S, Mirza, MA, Ma, J, Antwi-Afari, M F, et al. Multimodal integration for data-driven classification of mental fatigue during construction equipment operations: Incorporating electroencephalography, electrodermal activity, Developments in The Built Environment 2023 vol.15 pp. 2666-1659.

[6] Guo W. Zhao M. Alrashoud A. Tolba S. F and Xia F, Multimodal Educational Data Fusion for Students' Mental Health Detection, in IEEE Access,2022 vol. 10, pp. 70370-70382

[7] Joshi S, Jain K, Joshi U, Jain Y, Balpande S and Kulkarni R, MindMatrics: A Framework for Advanced Mental Health Diagnosis Using Multimodal Data Analysis, 2025 International Conference on Advanced Computing Technologies (ICoACT), Sivalasi, India, 2025, pp. 01-08

[8] Titung R, Interactive Machine Learning for Multimodal Affective Computing, 2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Nara, Japan, 2022, pp. 1-4

[9] Fusion K S et al, Net: A Multi-Modal Deep Learning Framework for Emotion Recognition and Mental Health Assessment, 2024 International Conference on Emerging Research in Computational Science (ICERCS), Coimbatore, India, 2024, pp. 1-5.

[10] Dalirinia E, Yaghoobi M, Tabatabaee H, et al. Multimodal Lotus Effect Algorithm for Engineering Optimization Problems. Engineering Reports, 2025, 7(4): e70137-e70137.