Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.20260307316

A Review of Multimodal Affective Computing in the Field of Mental Health Monitoring

Shunyao Zhang*

Beijing University of Posts and Telecommunications, Beijing, China

**Corresponding author: Shunyao Zhang.*

Abstract

Mental health is of great significance for an individual's development. As the worldwide mental health crisis grows more severe, depression and anxiety have turned into the primary causes of disability. Traditional approaches to evaluating mental health mainly depend on patients' self-reports and clinical interviews, which are frequently limited by strong subjectivity, high latency and social stigma, resulting in missed diagnoses. To overcome these limitations, multimodal affective computing integrates multi-source data such as text, speech, vision, and physiological signals to provide a revolutionary technical means for the objective quantification and monitoring of psychological states in real-time. This review systematically summarizes the technological advancements in this field, covering multimodal data processing, feature extraction, fusion architectures, and emotion recognition models. It also explores various application scenarios such as depression monitoring, stress management, and crisis intervention. Nevertheless, its application also encounters some difficulties, including technical problems, worries about data privacy and ethical dilemmas. Corresponding strategies involving technological invention, privacy-protecting computing, and ethical frameworks are discussed. Finally, this review concludes that multimodal affective computing has potential to transform mental health care. Its achievement depends on continued technological refinement and responsible, human-centered implementation within interdisciplinary collaboration.

Keywords

multimodal affective computing, mental health, depression monitoring, multimodal data processing

1. Introduction

Mental health problems have turned into one of the most pressing global public health challenges. Based on the latest statistics disclosed by the World Health Organization (WHO) and related institutions in 2025, more than one billion individuals across the world are suffering from various degrees of mental disorders, among which depression and anxiety disorders are the most common diseases [1, 2]. And there was also a distinct tendency for these problems to be more widespread among the young generation [2]. But the traditional distribution of medical resources is extremely imbalance, and in low-income nations, the great majority of patients cannot get basic treatment [1].

These limitations gave rise to multimodal affective computing and this technology aimed to attain objective, real-time, and non-intrusive monitoring of emotional states through the comprehensive analysis of an individual's text, speech, vision, and physiological signals [4, 5]. Studies [6, 7] have shown that multimodal systems possess considerably greater accuracy when it comes to detecting early indications of mental health crises in contrast to single-modal approaches and this review aims to summarize the latest technological advancements in this field, analyze the difficulties in its clinical applications, and discuss strategies for addressing these issues by leveraging cutting-edge technologies so as to offer theoretical support for the growth of developed and reliable mental health monitoring systems.

2. Current Research Status and Technical Evolution

In the area of emotion recognition and psychological state detection, machine learning and deep learning models constitute the technological core. Traditional machine learning models like Support Vector Machines (SVM) [14] and Random Forests [10], were extensively applied during the initial phases for preliminary sorting, valued for their interpretability and high computational efficiency when dealing with small-scale datasets [3].

As multimodal data grew more complex, deep learning models exhibited enhanced representation learning abilities. And recurrent Neural Networks (RNN), along with their variant Long Short-Term Memory networks (LSTM) [16] are skilled at processing temporally dependent data like speech audio and video frame sequences while Convolutional Neural Networks (CNN) [15] concentrates on extracting local features from spatial or spectral data [7].

Over the past few years, the Transformer architecture [17] has witnessed significant advancement in multimodal affective computing because of its self-attention mechanism, which can effectively model long-range connections and achieve deep interaction and alignment of cross-modal features [12].

Moreover, multi-task learning models [11] enhanced the generalization capacity and computational efficiency through the sharing of feature representations and optimization of several associated tasks such as emotion recognition and inference of psychological states [4].

End-to-end models, through end-to-end learning, can directly mine deep patterns from raw or shallowly processed multimodal data for emotion and psychological state discrimination [13].

3. Application Scenarios and Case Studies in Mental Health Monitoring

3.1 Early Recognition and Longitudinal Monitoring of Depression

In early screening for depression and anxiety, multimodal affective computing constructs risk assessment models by integrating individuals' behavioral, linguistic and physiological signals.

These models are typically based on multi-source data such as video interviews, voice recordings, and social media text. For instance, in the case of interview videos, computer vision algorithms are able to extract intensity and duration sequences of facial action units. Whose features are linked to the emotional numbness of depression or the anxious state [8], and at the same time, speech signal processing methods can analyze the prosodic characteristics of speech, such as decreased pitch variation range, a slower speech rate, and more frequent pauses. And these acoustic signs have been verified as potential biomarkers of depressive states [4], while text data from social media offered hints regarding psychological states in daily contexts via the natural language processing analysis of sentiment tendency, self-referencing frequency, and syntactic complexity [9].

Integrating these diverse-type modal data is a crucial technological aspect, and advanced models utilize decision-level or model-level fusion founded on deep learning to allow the model to autonomously learn complementary associations among different modal signals. A typical application is building an end-to-end screening system [18], which receives an interview video of a patient, automatically analyzes the absence of pleasure in facial expressions, monotony in speech, and negative inclinations in speech content, ultimately producing a risk probability score for depression or anxiety, and this analysis based on behavioral signals reduces complete reliance on patients' subjective self-reports.

Compared to frequently utilized traditional tools like PHQ-9, multimodal assisted screening demonstrates distinct benefits, because traditional scales depend on patients' immediate reporting by themselves, which can be influenced by memory bias, social desirability, and lack of understanding about the illness [3]. While multimodal models can pick up unconscious or subtle behavioral patterns, providing more objective quantitative evidence. Besides, scale assessments are discrete, whereas multimodal technology theoretically enables inexpensive, non-intrusive continuous monitoring.

3.2 Real-time Perception of Anxiety and Stress Management

Real-time stress and emotional state monitoring represents one of the most promising directions for multimodal affective computation in mental health. Its core lies in combining physiological signals from wearable devices with contextual information from smartphones so as to achieve continuous, objective evaluation of an individual's mental state.

Physiological signals like Heart Rate Variability (HRV) can directly mirror the activation degree of the autonomic nervous system and are reliable indicators of stress and emotional responses [19], for instance, a decrease in HRV is often associated with higher stress and weaker emotional regulation ability. Smartphones offer abundant contextual data, such as app usage frequency, communication styles, location changes, and physical activity amounts [8]. These behavioral patterns have connections with an individual's psychological state.

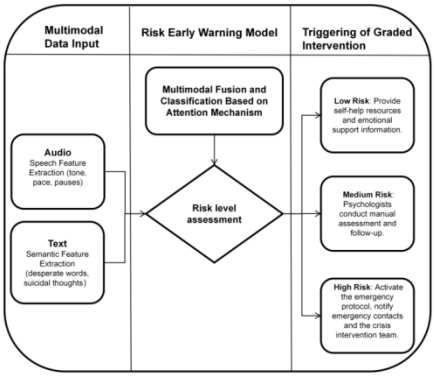
Through the fusion of these two kinds of heterogeneous data by means of machine learning models, personalized real-time monitoring systems can be constructed. The system initially carries out noise reduction and feature extraction on raw physiological signals while extracting behavioral features from phone logs. Then, feature-level or decision-level fusion strategies are employed; for example, models based on attention mechanisms can dynamically weigh the importance of physiological signals and contextual information in different situations [23]. The trained model can analyze the input data stream continuously. When characteristic patterns consistent with high pressure or negative emotions are detected, the system can trigger instant interventions. In workplace health management situations, such a system can be incorporated into health platforms. When long-term high-stress conditions are found in employees, it can automatically send mindfulness breathing instructions or suggest breaks. For daily emotion regulation, individual users can receive personalized suggestions via mobile apps according to their current physiological and behavioral state [20].

3.3 Suicide Risk Prediction and Crisis Intervention

In suicide risk alerting, as shown in Figure 1, multimodal affective computing constructs more sensitive and robust risk assessment models through the integration of data from diverse sources. Voice signals from emergency psychological hotline are filled with emotional hints, such as flat or rapid tone, irregular alteration in speech rate, and long silences. These acoustic features are strongly associated with negative emotions such as hopelessness and anxiety [4]. Simultaneously, text posted by individuals on platforms like statements of being worthless, feeling isolated, or explicit suicidal intentions, offers direct semantic-level danger signals [9]. To effectively fuse this heterogeneous information, the framework employs an attention-based mechanism.

Moreover, the model needs to learn patterns associated with suicide risk from the fused multimodal features. By training on large-scale, professionally annotated crisis intervention datasets, models can detect subtle evolutionary trajectories from general psychological distress to high-risk crisis. Once the model assesses the individual’s risk level, the system can trigger a graded intervention protocol, which is detailed in the right panel of Figure 1.

Figure 1: Suicide Warning and Intervention Framework Based on Multimodal Emotional Computing



4. Challenges in Application

4.1 Technical Challenges

When applying multimodal affective computing technology for mental health surveillance, we have to face technical hurdles across the entire pipeline.

During data collection and preprocessing, time synchronization and alignment of multimodal data are major barriers. Because emotional expression is a dynamic temporal process. Merely forcing alignment results in the loss of critical temporal dynamic information, but complicated asynchronous fusion models impose high demands on algorithm. The accuracy of this alignment affects the reliability of subsequent feature fusion and emotional state inference [22].

The cross-modal semantic gap represents another challenge as different modal data exist within heterogeneous feature spaces, and there are substantial differences in the granularity, abstraction level, and reliability of their emotional representations. How to construct a unified semantic space where features from different modalities can interact and complement, rather than simply being concatenated, is key to model design. Some studies [18, 23] attempt to model complex inter-modal relationships by introducing attention mechanisms or graph neural networks. However, enabling models to comprehend the cooperation and opposition between “a frowning micro-expression” and “pauses and tremors in speech” within the emotional state of “anxiety”, deeper semantic alignment and reasoning systems are still necessary.

During training, a small number of samples and serious class imbalance restrict model training and performance [24]. Obtaining high-quality multimodal mental health datasets is expensive and requires strict ethical review, which leads to limited data scale. Meanwhile, the distribution of mental health states is imbalanced as there are far fewer samples of high-risk states than normal states. This made it easy for models to be biased towards the majority class, resulting in bad recognition performance [3]. Although some research has employed data augmentation, cost-sensitive learning, or transfer learning to mitigate this problem, data scarcity still remains a significant barrier to model generalization in cross-domain and cross-population applications.

A major stumbling block for implementation is the lack of model generalization capability as data collected in laboratory environments is mainly obtained in controlled, quiet, low-interference scenarios where subjects know they’re being observed, which may lead to performance bias [25]. But real-world mental health monitoring environments are complicated and changeable with diverse interferences. Moreover, people have distinctions in cultural background, expressive habits, and physiological baselines [26]. Thus, models that perform well when trained on one group may suffer severe performance decline when applied to another group. Therefore, developing robust models is a necessary gap to bridge for moving technology from the laboratory to practical application.

4.2 Data and Privacy Challenges

Mental health data presents data and privacy challenges because of its sensitivity. Acquiring high-quality, large-scale labeled data is a primary challenge. Emotion labeling is subjective, and labeling mental health conditions demands clinical knowledge, which renders labeled data meeting research needs scarce and costly. The multimodal nature of data further complicates labeling and any labeling inconsistency can impact the reliability of subsequent model training.

There are also substantial risks of privacy leakage at every stage of the data lifecycle. In the data collection phase, multimodal data utilized for emotion analysis may expose users' identity information details, locations, social connections, and even undiagnosed mental illnesses [27]. Then during storage and transmission, centralized multimodal databases, if attacked, could lead to irreversible privacy harm [29].

Furthermore, strict laws, regulations, and ethical evaluations form another level of constraint. For example, frameworks like the European Union's General Data Protection Regulation (GDPR) and the United States' Health Insurance Portability and Accountability Act (HIPAA) set more demanding standards for handling special type of data like mental health in terms of consent, security, and transparency [29]. So, researchers must incorporate privacy protection into their technical process. These compliance requirements increase the complexity of technical implementation, but they are the cornerstone for promoting the healthy, sustainable development of research in this field and ultimately gaining user trust.

4.3 Ethical and Social Acceptance Challenges

Multimodal affective computing algorithms may amplify social biases. If training data cannot balance to cover diverse demographic groups, models are likely to have performance differences when identifying the emotional states of people from specific races, genders, or cultural backgrounds, and this technical discrimination could result in wrong judgment or overlooking the mental health conditions of minority groups, thus aggravating health inequalities. The black box characteristic of algorithms makes it tough to trace and correct bias. Without careful data management and algorithmic examination, technological tools themselves might turn into means for deepening social bias [5, 30].

Besides, excessive technological penetration into the private realm of emotion may influence interpersonal alienation and provoke the risk of individual stigmatization. When emotional states are continuously quantified into monitorable metrics, the natural compassion in interpersonal interactions may be replaced by cold algorithm scores. Moreover, the diagnostic labels like "depression tendency" output by system may generate psychological hints for users and lead to stigma in social settings, preventing them from taking the initiative to seek assistance [3].

5. Coping Strategies and Frontier Progress

5.1 Technological Strategies

To deal with technical challenges like insufficient data, modality heterogeneity, and noise interference, a series of technological strategies have been proposed and are developing. Regarding the problem of getting high-quality labeled data, self-supervised learning and transfer learning have become useful solutions as self-supervised learning is able to learn general representations from a large amount of unlabeled multimodal data by designing proxy tasks, providing high-quality features for subsequent emotion recognition tasks [13]. And transfer learning use large models pre-trained in general areas to transfer their knowledge to the specific mental health domain, greatly decreasing the reliance on the amount of labeled data [4].

At the feature fusion level, the key step is to design effective cross-modal attention systems, as traditional simple combination or early fusion methods have difficulty in capturing complex nonlinear interactions among different modalities [18]. Therefore, researchers have put forward multi-level, fine-grained fusion architectures. One strategy is to use a two-stage attention mechanism that dynamically allocates weights at the temporal step and modality levels, thus achieving more precise feature fusion. Another innovation is introducing a dual-level gated segment fusion model [31], which enhances cross-modal interaction strategies and integrates local-global features to make full use of emotional detail information. Additionally, employing

a self-attention mechanism based on loop generation to achieve feature conversion between modalities, thus strengthening model robustness when some modalities are absent [32].

5.2 Privacy Protection and Data Security Strategies

Privacy safeguarding and data security serve as the cornerstones for technology application implementation [3]. Thus, a series of technical and management strategies have been proposed and applied. Federated Learning allows models to be trained on local devices, with only encrypted model updates uploaded to a central server for aggregation, avoiding concentrated exposure of sensitive data at the source. It is particularly suitable for affective computing tasks involving multimodal data, enabling the learning of effective global models from decentralized data while protecting data privacy [28]. Complementary to Federated Learning is Homomorphic Encryption technology [33]. It allows computations on encrypted data, with the decrypted results matching those of computations on plaintext data, which means servers can perform aggregation operations without decrypting users uploaded encrypted data, providing stronger end-to-end security for federated learning or other cloud computations.

What's more, Edge Computing architecture considerably reduces the amount of data needing upload to the cloud by shifting data processing and model inference tasks to terminal devices or edge servers close to the data source [34]. In mental health monitoring scenarios, this means users' raw audio, video, or physiological signals can undergo preliminary feature extraction or even emotional state inference locally, and only necessary, anonymized intermediate results or final analysis conclusions are uploaded, which effectively reduces the risk of data leakage during transmission and cloud storage.

Besides the core technologies mentioned above, data minimization and anonymization are fundamental principles across the whole data lifecycle [27]. These technical and management solutions together form a multi-layered, in-depth privacy protection system, providing a feasible deployment path for multimodal affective computing in the sensitive field of mental health.

5.3 Ethical Frameworks

As multimodal affective computing technology is increasingly applied in mental health monitoring, it is crucial to establish responsible ethical frameworks and standardization systems [3]. The key of this process lies in designing and carrying out comprehensive AI application ethical guidelines, and algorithm fairness is a main concern. Meanwhile, model interpretability is the cornerstone for the trust between users and clinicians. Complex deep learning models are often seen as black boxes, with their decision logic difficult to trace. In the sensitive field of mental health, applying XAI methods to reveal the key rationale behind model decisions can not only assist professionals in verifying result reliability but also provide a clear path for manual review in case of conflicts and avoid blind reliance on conclusions from the model [35].

Constructing effective human-AI collaborative decision-making mechanisms is crucial for striking a balance between technological efficiency and humanistic concern [36], so multimodal affective computing systems should be positioned as auxiliary tools instead of final decision makers; their results need to be combined with comprehensive assessment operated by clinicians. System design must clarify the boundary of human and AI responsibilities. This collaborative mechanism can improve assessment efficiency and objectivity while preserving professionals' ultimate discretion and empathy in complex situations [37], finally enabling multimodal affective computing towards trustworthy, and responsible application in the field of mental health.

6. Discussion - Future Development Trends

The future development of multimodal affective computing will revolve around dimensions such as data acquisition, emotional comprehension, intervention cycles, and model customization. Data acquisition is trending towards being more natural and less intrusive to reduce disturbance to the monitored individual. Based on non-contact sensors like RGB-D cameras, combined with wearable devices and smart environment sensing technologies, it will become possible to capture multi-dimensional data. This non-intrusive mode not only enhances data continuity and authenticity but also establishes the foundation for long-term, regular mental health monitoring.

Moreover, the mission of affective computing is gradually shifting from simple emotion state recognition to deeper understanding of emotions and empathy. Future models need to interpret the context of emotional generation, intensity dynamics, and mixed emotional states rather than simply classifying discrete emotion labels. This requires advanced algorithms to fuse contextual semantics, and by introducing advanced attention mechanisms, models can better grasp complex inter-modal relations, achieving a leap from recognizing emotions to truly understanding mood.

Besides, the deep combination with Digital Therapeutics, will support the construction of comprehensive monitoring-analysis-intervention systems. Multimodal affective computing can function as the core sensing module of Digital Therapeutics and assess users' psychological state changes in real time. When the system detects significant signs of depression, anxiety, or stress, it can automatically set off or suggest personalized intervention content, making mental health support from passive reaction to active prevention and immediate adjustment, significantly improving the timeliness and effectiveness of interventions.

Finally, developing personalized and adaptive models is crucial for dealing with individual differences. Future systems need to dynamically adjust feature extraction and fusion strategies based on different user information like behavior patterns, physiological baselines, and historical data. Through continuous learning and model fine-tuning, systems can keep adapting to changes in users' emotional expression habits, thus boosting the accuracy and reliability of long-term monitoring. Ultimately, the combination of these technologies will push mental health monitoring towards precision, intelligence, and humanization, offering robust technological support for achieving inclusive mental health services.

7. Conclusion

Multimodal affective computing opens up a more comprehensive, objective, and dynamic path for assessing mental health monitoring. Its core lies in breaking through the limitations of traditional single modalities or subjective scales so as to achieve more subtle and continuous perception and interpretation of individual emotional states. The current research shows that, technically speaking, development has progressed from early feature fusion to deep fusion, with a series of advanced model architectures proposed to optimize this process. These advancements converge towards one shared goal: improving the accuracy and robustness of emotion recognition, laying the foundation for precise mental state analysis.

At the application level, multimodal systems are moving from laboratory to practical scenarios. Research has begun to construct AI-powered monitoring systems that integrate multimodal data, indicating that it might be possible to achieve full-chain services from screening and warning to intervention and feedback in the future. However, the field faces many challenges on its path to mature application. These challenges include not only technical aspects, but also profound ethical, privacy problems. The sensitivity of data and the need for explainable fairness require us to set up strict standards and norms.

Therefore, optimizing algorithms alone cannot lead to the ultimate success of multimodal affective computing in the mental health field. It is in urgent need of profound combinations and collaborative endeavors from domains like computational science, psychology, clinical medicine, and ethics. Future research needs to push forward the technological boundary while placing human value at the center, truly empowering mental health and well-being through technology by building trustworthy, reliable, and human-centered intelligent systems.

References

[1] World Health Organization, World mental health today: latest data. Geneva: World Health Organization, 2025. [Online]. Available: https://iris.who.int/

[2] L. Taylor, “One billion people have mental health conditions, WHO says,” BMJ, vol. 390, p. r1860, Sep. 2025. doi: 10.1136/bmj.r1860

[3] P. Cruz-Gonzalez et al., “Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications,” Psychological Medicine, vol. 55, no. 3, p. e18, Feb. 2025. doi: 10.1017/S0033291724003295

[4] Y.-H. Hu, R.-Y. Wu, M.-Y. Su, I.-L. Lin, and C.-C. Shen, "Multimodal Multitask Learning for Predicting Depression Severity and Suicide Risk Using Pretrained Audio and Text Embeddings: Methodology Development and Application," JMIR Mental Health, vol. 11, p. e53457, Jan. 2024. doi: 10.2196/53457

[5] M. Schlicher, Y. Li, S. M. K. Murthy, Q. Sun, and B. W. Schuller, "Emotionally adaptive support: a narrative review of affective computing for mental health," npj Digital Medicine, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12568696/

[6] Y. Wu et al., "A Comprehensive Review of Multimodal Emotion Recognition," Biomimetics, vol. 10, no. 7, p. 418, 2025. doi: 10.3390/biomimetics10070418

[7] D. Mamieva et al., "Attention-Based Approach to Multimodal Emotion Recognition," Sensors, vol. 23, no. 12, p. 5475, Jun. 2023. doi: 10.3390/s23125475

[8] J. Chen et al., "Multimodal digital assessment of depression: integration of actigraphy and a mobile app," Translational Psychiatry, vol. 14, no. 1, p. 154, Mar. 2024. doi: 10.1038/s41398-024-02873-4

[9] R. Huang et al., "Exploring the Role of First-Person Singular Pronouns in Detecting Suicidal Ideation: A Machine Learning Analysis of Clinical Transcripts," Diagnostics, vol. 14, no. 3, p. 225, Jan. 2024. doi: 10.3390/diagnostics14030225

[10] C. Á. Casado et al., "Depression Recognition Using Remote Photoplethysmography From Facial Videos," IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 3125-3137, Oct. 2023. doi: 10.1109/TAFFC.2023.3274567

[11] M. Zhao et al., "Decoupled Multi-Perspective Fusion for Speech Depression Detection," IEEE Transactions on Affective Computing, vol. 16, no. 3, pp. 1772-1786, Jul.-Sep. 2025. doi: 10.1109/TAFFC.2025.3538519

[12] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021. doi: 10.48550/arXiv.2010.11929

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 12449-12460, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[14] N. Janardhan et al., "Improving Depression Prediction Accuracy Using Fisher Score-Based Feature Selection and Dynamic Ensemble Selection Approach Based on Acoustic Features of Speech," Traitement du Signal, vol. 39, no. 1, pp. 87-107, Feb. 2022. doi: 10.18280/ts.390109.

[15] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," Sensors, vol. 21, no. 9, p. 3046, May 2021. doi: 10.3390/s21093046.

[16] A. Shenoy and A. Sardana, "Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 19-30. doi: 10.18653/v1/2020.acl-main.3.

[17] S. Siriwardhana, A. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion," IEEE Access, vol. 8, pp. 176274-176285, Sep. 2020. doi: 10.1109/ACCESS.2020.3026823.

[18] N. K. Iyortsuun, S.-H. Kim, H.-J. Yang, S.-W. Kim, and M. Jhon, "Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features," IEEE Access, vol. 12, pp. 20479-20489, 2024. doi: 10.1109/ACCESS.2024.3361111

[19] G. J. Martinez et al., "Alignment Between Heart Rate Variability From Fitness Trackers and Perceived Stress: Perspectives From a Large-Scale In Situ Longitudinal Study of Information Workers," JMIR Mhealth Uhealth, vol. 10, no. 3, p. e33754, Mar. 2022. doi: 10.2196/33754

[20] W. Huang et al., "Mobile apps, AI, and teletherapy: a comprehensive review of digital mental health tools for nurse," Frontiers in Public Health, vol. 13, 2025. doi: 10.3389/fpubh.2025.1686766

[21] S. Jere and A. P. Patil, "Detection of Suicidal Ideation Based on Relational Graph Attention Network with DNN Classifier," International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 10s, pp. 321–332, Jul. 2023. [Online]. Available: https://www.ijisae.org/index.php/IJISAE/article/view/3255

[22] R. Xiao, C. Ding, and X. Hu, "Time Synchronization of Multimodal Physiological Signals through Alignment of Common Signal Types and Its Technical Considerations in Digital Health," Journal of Imaging, vol. 8, no. 5, p. 120, 2022. doi: 10.3390/jimaging8050120.

[23] K. Zhao et al., "Multimodal Sentiment Analysis—A Comprehensive Survey From a Fusion Methods Perspective," IEEE Access, vol. 11, pp. 91321-91345, 2023. doi: 10.1109/ACCESS.2023.3308316.

[24] Y. Li et al., "Automated Depression Detection From Text and Audio: A Systematic Review," IEEE Journal of Biomedical and Health Informatics, May 2025. doi: 10.1109/JBHI.2025.3570900.

[25]M. Richter et al., "Generalizability of clinical prediction models in mental health," Molecular Psychiatry, 2025.

[26] V. Farsadaki et al., "AI affective computing and behavioral health," Frontiers in Computer Science, vol. 7, 2025. doi: 10.3389/fcomp.2025.1692728.

[27] A. Mandal et al., "Towards Privacy-aware Mental Health AI Models: Advances, Challenges, and Opportunities," arXiv preprint arXiv:2502.00451, Feb. 2025. [Online]. Available: https://arxiv.org/abs/2502.00451

[28] P. Dubey et al., "Federated learning for privacy-enhanced mental health prediction with multimodal data integration," Computer Methods in Biomechanics and Biomedical Engineering, 2025. doi: 10.1080/21681163.2025.2509672

[29] S. T. Shah et al., "Federated Learning in Public Health: A Systematic Review of Decentralized, Equitable, and Secure Disease Prevention Approaches," International Journal of Environmental Research and Public Health, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12607528/

[30] P. Singhal et al., "Domain adaptation for bias mitigation in affective computing: use cases for facial emotion recognition and sentiment analysis systems," Discover Applied Sciences, vol. 7, p. 229, 2025. doi: 10.1007/s42452-025-06659-1

[31] J. Huang et al., "Multimodal alignment and hierarchical fusion network for multimodal sentiment analysis," Electronics, vol. 14, no. 19, p. 3828, 2024. doi: 10.3390/electronics14193828.

[32] Z. Liu et al., "Intelligent assessment of English teachers' classroom language interaction and emotional behaviour based on artificial intelligence," Sci Rep, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12528482/

[33] E. A. Mantey et al., "Federated Learning Approach for Secured Medical Recommendation in Internet of Medical Things Using Homomorphic Encryption," IEEE Journal of Biomedical and Health Informatics, Jun. 2024. doi: 10.1109/JBHI.2024.3411032

[34] L. Xia et al., "Enabling University Mental Health Monitoring Through 6G Edge-Cloud IoT Environmental Perception Frameworks," IEEE Communications Standards Magazine, Jan. 2025. doi: 10.1109/MCOMSTD.2024.00045

[35] S. Hameed et al., "Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models," Frontiers in Artificial Intelligence, vol. 7, 2025. doi: 10.3389/frai.2025.1627078

[36] S. J. M. Smith et al., "AI–Supported Shared Decision-Making (AI-SDM): Conceptual Framework," Journal of Medical Internet Research, vol. 27, p. e75866, Jan. 2025. doi: 10.2196/75866

[37] A. Kerasidou, "Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare," Bulletin of the World Health Organization, 2020.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.