AI-Based Emotion State Monitoring: Methods, Applications, and Limitations

Chen Ziqi
Data Science and Big Data Technology, North China University of Technology, NCUT, Beijing, 100144, China
2077673769@qq.com

ABSTRACT

Emotional state monitoring has become an important research topic of artificial intelligence, which has applications in mental health, human-computer interaction and intelligent systems. This paper provides a problem-oriented review of emotion monitoring technology based on AI. Firstly, the basic challenges of emotion recognition are analyzed, including subjectivity, context dependence and the complexity of multi-mode. Next, the main application scenarios, such as mental health monitoring and intelligent interaction, are reviewed. Finally, the limitations of current methods-including data quality, interpretability, privacy and moral hazard-are critically analyzed, and the future research direction is put forward. This review shows that the current emotion recognition system is still mainly based on probabilistic reasoning, rather than real cognitive understanding, emphasizing the need to balance technological progress and moral responsibility.

Keywords: Emotion Recognition; Affective Computing; Multimodal Fusion; Large Language Models; Mental Health Monitoring; Human-Computer Interaction; Explainable AI; AI Ethics.

1. INTRODUCTION

With the rapid development of artificial intelligence, emotion recognition has become the core theme of emotion computing. This enables machines to perceive, understand and respond to human emotional state, thus improving the quality of interaction and contributing to applications such as mental health monitoring and personalized service [1].

However, emotion recognition is very different from the traditional data analysis task. Emotional states are subjective, constantly changing and highly dependent on context, which makes it difficult to model them accurately.

This paper aims to provide an organized overview of AI-based emotion monitoring technology, focusing on changing methods, application scenarios and major challenges.

2. CORE CHALLENGES AND METHODOLOGICAL EVOLUTION

2.1 Fundamental Challenges in Emotion Monitoring

In this paper, emotion recognition and emotion monitoring are similar but not identical ideas. Emotional recognition is a process of identifying or classifying a person’s emotional state according to observable signs (such as text, voice, facial expression or body data). Emotional monitoring, on the contrary, is to observe the emotional state continuously or repeatedly for a period of time. It is usually used in practical applications, such as evaluating mental health, human-computer interaction, or adapting to intelligent systems. Therefore, emotion recognition can be considered as an important technical part of emotion monitoring, which represents a broader application-oriented framework.

Emotions cannot be seen directly. They must be guessed from indirect signs such as text, speech, facial expressions and physiological data. Each of these ways of doing things has limitations, for example, they can be sensitive to noise, ambiguous, or change from one person to another [1, 2].

The function-component-representation framework suggests that affective computing asks three fundamental questions: why to compute the affect, how to compute the affect, and what representation of the affect should be computed [1]. This shows that emotion recognition is not just a technical problem, but also a conceptual issue related to how we understand emotion itself.

2.2 Unimodal Monitoring Signals: Trade-offs Between Accessibility and Expressiveness

Early emotional monitoring systems usually used only one signal, such as voice, text, facial expression or body data. Traditional machine learning methods, such as SVM and KNN, need manual features and are difficult to understand complex emotions [3].

Emotional recognition in speech analyzes things such as tone, height and rhythm. However, it is not always effective because of different languages, surrounding noise or different ways people speak [3]. Deep learning is very effective in recognizing facial emotions, but if you hide the face, the light changes or the facial movements are very small, problems will occur [4].

Body signals, such as ECG, PPG and EEG, can provide more objective information about emotions. Electrocardiogram (ECG) and PPG can show the reaction of nervous system, and EEG can see the brain activity related to emotion [5,6]. But for these methods, special sensors are needed, and the results may vary with people or signal noise [6].

Therefore, there is a big problem with methods that use only one data type: either they are easy to use or they are very accurate, but they cannot be used at the same time.

2.3 Multimodal Fusion in Emotion Monitoring: From Complementarity to Complexity

In order to overcome the limitations of single-mode method, multimodal emotion recognition combines multiple data sources. Combining speech, facial expression, text and physiological signals can improve robustness and accuracy by using complementary information between patterns [7,8].

In emotional monitoring, multi-mode fusion is very important, because the emotional state is dynamic and cannot be reliably captured by a single signal. Recent models, such as VAF-Net and MURBN, show how multi-pattern recognition module can help to create a more powerful emotion monitoring system.

VAF-Net is a lightweight model, which is used to identify emotions with multiple patterns. It is designed to be more accurate and use less calculation. Its core mechanism is not simply the audio-visual fusion model, but the combination of facial image features, voice information and text mode embedding through the integration of multi-mode features. Specifically, Zhang et al. [9] used ResNet-50 as the main extractor of facial images, and adopted the multi-modal attention fusion (MAFM) mechanism to integrate different patterns of information. The model also used the mixing and refining of knowledge to become lighter, making it more suitable for practical deployment in the case of scarce resources, such as robot emotion monitoring.

Murbn (multimodal unified recurrent-bottle neck network) is another representative model of multimodal emotion recognition. Hu et al. [10] propose a two-stage strategy of “encoding and alignment first, fusion and refinement later.” In this model, the Recurrent Unimodal Modulation Network (RUMN) contextually encodes temporal features from different modalities and reduces modality heterogeneity, while the Asymmetric Bottleneck Fusion Network (ABFN) enables efficient and low-noise cross-modal information interaction and fusion. Compared with simply concatenating multimodal features, MURBN focuses more on feature alignment, redundancy reduction, and robust cross-modal representation learning.

However, multimodal systems also introduce new challenges, including modality heterogeneity, synchronization difficulty, data alignment problems, and high computational cost [7,8]. Therefore, the core problem shifts from insufficient information to complex information integration.

2.4 Large Model Paradigm: Context-Aware Emotion Monitoring and Its Limits

Large-scale language models and multimodal converters expand emotional monitoring. In the past, people classified emotions one by one. Now, we can analyze the emotional state according to the context. These models can handle complex semantic information. They help to analyze the emotions in the AI dialogue system [12].

2.5 Key Recognition Techniques Supporting Emotion Monitoring

The deep learning model has greatly improved the system of emotion recognition. They allow automatic extraction of functions and learning of representations. Among them, Convolutional Neural Network (CNNs), Recurrent Neural Network (RNNs) and Transformer-based model play a core role.

CNN is widely used to extract spatial features from visual data and spectrograms. They can automatically learn hierarchical representations, from simple edges and textures to complex semantic patterns. For facial emotion recognition, CNN can capture facial muscle movement and expression changes well. The hybrid method combines the manually created functions, such as local binary pattern (LBP), gradient-oriented histogram (HOG) and CNN architecture, and shows better robustness under different conditions [20].

In speech emotion recognition, CNN with spectrogram is used. It can help the model find emotional clues, such as the change of sound height and energy distribution. Research shows that the model based on CNN can perform well by learning directly from the original features of audio [21].

RNN and its variants, especially long-term and short-term memory (LSTM) networks, are used to understand the time dependence of sequential data. Because the emotional expression changes with time, LSTM network can capture this dynamic well and improve the recognition performance [22].

The mixed architecture of CNN and LSTM has become a very important method in speech emotion recognition. CNN layer extracts features, and BiLSTM layer captures dependencies over time. It allows high accuracy to be obtained on actual data sets [23].

Recently, the model based on Transformer introduced attention mechanism into emotion recognition task. Based on the architecture of Visual Transducer (ViT), great progress has been made in facial emotion recognition performance through the enhancement of attention mechanism [24].

2.6 Multimodal Fusion Strategies

Multi-modal emotion recognition relies on the integration of information from multiple data sources, such as voice, facial expression and physiological signals. The effectiveness of these systems depends on the fusion strategy used.

Early fusion combines the characteristics of different modes at the input level, allowing cross-mode interaction, but increasing the size complexity. Post-fusion deals with each mode independently, and combines the results in the decision-making stage to provide flexibility, but cross-mode learning is weak.

These two strategies are mixed and integrated, which is the most widely used method at present; For example, the wearable assisted multimodal system combines EEG, ECG and facial signals using a hybrid deep learning framework, which greatly improves the performance of stress and nervous system disease monitoring [25].

2.7 Benchmark Datasets for Emotion Recognition

The performance of emotion recognition model depends largely on the availability and quality of benchmark data sets. Several widely used data sets have been created to help the research in this field.

DEAP data set is one of the most commonly used data sets in multimodal sentiment analysis. It includes electroencephalogram (EEG), peripheral physiological signals and video recordings collected from participants watching emotional stimuli. This data set provides discrete emotion labels and continuous evaluation, including benchmark, price and dominance, making it suitable for classification and regression tasks.

Another important database is MAHNOB-HCI. It contains synchronous multimodal data, such as facial video, audio signal, EEG recording and gaze information. This database is widely used to study multimodal emotion recognition and human-computer interaction scenarios.

IEMOCAP data set is widely used in speech emotion recognition. It includes emotional speech data, playback and spontaneity, as well as motion capture and video recording. This data set allows emotions to be classified into multiple categories and has become a standard for evaluating speech-based models.

Although they are important, the existing data sets still have limitations. For example, they often have a small number of samples, the experimental conditions are highly controlled, and they lack the diversity of the real world. These limitations reduce the ability of the training model to run well in the new situation. This shows that we need more real and larger data sets for future research.

3. APPLICATIONS OF EMOTION MONITORING

3.1 Mental Health Monitoring

Emotion recognition based on artificial intelligence is widely used to monitor mental health. Machine learning model can predict postpartum depression by analyzing behavioral and physical data, so that risks can be seen early and actions can be taken [14].

The method based on Transformer also allows us to detect depression, identify negative emotions, emotional withdrawal and cognitive distortion from texts on social media [15].

The multimodal system combining facial, voice and text data allows continuous and personalized monitoring, which is helpful for early diagnosis and planning strategies to help mental health [16].

3.2 (Extended) Empirical Studies in Applications

Recent research shows that emotion recognition works well in real life. In health care, multimodal systems with wearable devices are more accurate in detecting stress and nervous system diseases because they combine physiological and visual signals [25].

In human-computer interaction, the chat robot that understands emotions greatly improves users' interest and makes the system more credible. The experimental results show that compared with the traditional system, adding emotion recognition mode improves the quality of interaction with users [26].

3.3 Human-Computer Interaction

Emotion recognition enhances interactive systems by allowing appropriate responses based on user emotions. In conversational artificial intelligence, multimodal emotion recognition can integrate text, voice and visual signals to produce a more situational awareness and empathy response [12].

However, emotional sensitive systems may also have unexpected consequences, such as emotional dependence, excessive trust in AI system and decreased user autonomy [12,17]. Transparency issues arise when users don't know that emotional data is being collected and analyzed.

3.4 Intelligent Systems and Adaptive Applications

Systems for understanding emotions are widely used. They exist in recommendation systems, educational platforms, assistive technologies and adaptive software. By listening to users' emotions, these systems can change what they display and how they interact. This makes things more personal and interesting [17].

However, several things need to be done to make these systems work normally. Emotional recognition must be reliable. User consent emotional intervention is appropriate.

4. LIMITATIONS AND ETHICAL CHALLENGES

4.1 Data Quality and Generalization

Many data sets are made in controlled places, so they can't show how to express emotions in real life well [7,8]. Therefore, it is subjective to label emotions. This makes the model unreliable and ineffective in other cases.

4.2 Model Interpretability

Deep learning models usually look like black boxes. Although they may be very accurate, we don't fully understand how they make decisions. This is why Explainable AI is very important, so we have more confidence. This is especially true in the field of health and mental health [15].

4.3 Privacy and Ethical Risks

Emotional data are very sensitive because they show our inner feelings. If used improperly, it may lead to problems: our privacy is no longer respected, we may try to manipulate our emotions, and we may make unfair decisions, such as in work, school or business recommendation systems [18,19].

When emotion recognition is used to organize or manage the public, the risk is even greater. In this case, people often cannot really control their data collection.

4.4. Cultural and Contextual Limitations

Emotion recognition models may not work well in different cultural backgrounds, because people don't express and convey their emotions in the same way. China's emotion recognition problem shows that it is difficult for AI systems to understand emotions that are implicit, indirect or highly dependent on context [13].

This limitation tells us that the high accuracy of the model on the benchmark data set does not always mean that it is reliable in real social situations.

4.5 Gap Between Recognition and Understanding

A basic limitation is the gap between pattern recognition and real emotional understanding. The current model mainly relies on the statistical correlation between the input signal and the emotional state of the marker [1,13].

Although this method allows classification, it fails to grasp the deeper cognitive and psychological mechanism of emotions. Therefore, the emotion recognition system must be regarded as a prediction tool, rather than a system that can truly realize emotion cognition.

5. FUTURE DIRECTIONS

Future research should focus on five aspects. First of all, we need to build a large-scale and real emotional data set to improve generality. Secondly, we need a personalized emotion recognition model to adapt to individual differences. Thirdly, it is necessary to develop advanced multi-mode fusion methods to improve the alignment and integration between different data sources [7,8]. Fourth, introduce interpretable artificial intelligence to improve the transparency and trust of sensitive applications [15]. Finally, an ethical framework and regulatory policies must be established to prevent the abuse of emotional data [18,19].

6. CONCLUSION

Emotional state monitoring based on AI has developed rapidly. It has changed from the traditional single-mode method to a multi-mode system, and recently it has changed into a large-scale model paradigm. These advances have greatly improved the accuracy, robustness and applicability of emotion recognition technology in medical care, human-computer interaction and intelligent systems.

However, despite these technological advances, there is still an important limitation: the current system is still mainly based on statistical reasoning, rather than real emotional understanding. This gap between recognition and cognition shows that even a very accurate model may not be able to grasp the underlying psychological mechanism of human emotions, especially in complex and context-dependent situations.

According to this review, the future progress of emotion recognition should not be judged only by figures such as accuracy or F1 score. Instead, we should pay more attention to interpretability, adaptability and reliability in real life. Most importantly, when you combine multimodal data with large models, you need to better understand how human emotions work, not just optimize the data.

In addition, this paper says that we should develop emotion recognition technology by considering their influence on society more. As these systems appear more and more in our daily life, the risks related to privacy invasion, emotional manipulation and algorithmic bias cannot be ignored, so technological innovation must go hand in hand with moral reflection and regulatory framework to ensure responsible deployment.

In addition, from a practical point of view, we need to bridge the gap between laboratory research and real-world applications. Many existing models are trained and evaluated on controlled data sets, which may not accurately reflect the real interpersonal interaction. Future systems should give priority to robustness in dynamic environment and include customized modeling to better adapt to individual differences.

In a word, this study shows that the next step of emotion recognition research should be to develop towards a more people-centered paradigm. The future system should focus on understanding, explaining and coping with emotions in a way that conforms to human cognition and social norms, rather than just detecting emotional States. Only through this transformation can AI-based emotion monitoring technology achieve technological progress and social acceptance.

REFERENCES

[1] Ma, H., & Yarosh, S. (2023). A review of affective computing research based on function–component–representation framework. IEEE Transactions on Affective Computing, 14(2), 1655-1671

[2] Hu, H., & Sun, T. (2025). A review of multimodal emotion recognition technologies: Integrating speech, facial expressions, and physiological signals. In Proceedings of ICICML.

[3] Rani, R., & Ramaiya, M. K. (2023). Detection of emotions from speech using deep learning techniques and traditional techniques: A survey. In Proceedings of ICACRS.

[4] Ansari, S., Gurudas, V. R., Kulkarni, P., & Rajesh, T. M. (2023). Facial emotion detection using deep learning: A survey. In 2023 IEEE International Conference on Contemporary Computing and Communications (InC4) (pp. xxx–xxx). IEEE. https://doi.org/10.1109/InC457730.2023.10262904

[5] Huang, X., Peng, Y., Huang, Y., Yao, W., He, S., Shan, Q., & Xiang, X. (2024). Emotion recognition based on ECG and PPG features. Computer & Telecommunication, 0(8), 33-39. https://doi.org/10.15966/j.cnki.dnydx.2024.08.023

[6] Zhou, Z., Sun, T., Wu, C., Liang, Z., & Li, G. (2025). A review of EEG-based emotion recognition model s.Journal of Xiangtan University (Natural Science Edition), 47(3), 35-53. https://doi.org/10.13715/j.issn.2096-64 4X.20250205.0001

[7] Syafnur, A., Sunyoto, A., Setyanto, A., & Kusrini. (2025). A systematic literature review on deep learning based multimodal facial expression recognition: Trends, gaps, and future opportunities.

[8] Neeli, V. K., Ghatty, K. L., Toluchuri, S., Ponnekanti, N. S., & Babu, B. V. S. (2025). Multimodal emotion recognition using facial, speech, and unstructured data for personalized music and mental health monitoring. In Proceedings of ICMNWC.

[9] Zhang, C., Xi, Y., Mei, W., Qi, H., & Dai, Z. (2026). VAF-Net: A lightweight multimodal emotion recognition model. Science and Technology & Innovation, (5), 6-9, 18. https://doi.org/10.15913/j.cnki.kjycx.2026.05.002

[10] Hu, J., Chen, J., Long, Y., Tian, X., Qian, C., & Chen, Y. (2026). Multimodal emotion recognition algorithm based on a unified recurrent bottleneck network. Journal of Southwest Minzu University (Natural Science Edition), 52(2), 214-221. https://doi.org/10.26978/j.cnki.xnmdzk.2026.02.012

[11] Wu, A., Wang, H., Liu, L., & Shi, W. (2026). A survey on large models for emotion recognition. Journal of Computer Science and Exploration, 20(3), 625-649.

[12] Singh, A. K., Binu, A., Chauhan, J., & Dubey, P. (2025). Enhancing conversational AI with emotion recognition: A multimodal approach. In Proceedings of IEEE MPCON.

[13] Wu, X. (2025). The “DeepSeek Chinese emotion problem” and possible solutions: A co-presence-predictive AI approach. Science, Economy and Society, 43(5), 36-46. https://doi.org/10.19946/j.issn.1006-2815.2025.05.005

[14] Devi, S., Dumbre, D., Chavan, R., Kumar, R., & Sable, N. (2025). AI-driven prediction of postpartum depression: A machine learning approach for early risk assessment.

[15] Laxman, V. S., Kulkarni, S. S., Hareesh, B. V. N., & Enduri, M. K. (2024). Explainable depression detection in social media using transformer-based models: A comparative analysis of machine learning. In Proceedings of IEEE CICN.

[16] Charaan, R. M. D., Rahul, B., Reddy, M. P., Santhoshkumar, S. P., Udayakumar, K., & Sasikumar, M. (2025). Smart emotion recognition for personalized mental health support.

[17] Aranha, R. V., Corrêa, C. G., & Nunes, F. L. S. (2021). Adapting software with affective computing: A systematic review. IEEE Transactions on Affective Computing, 12(4), 883-899.

[18] Ruan, K. (2023). Problems, risks, and governance of emotion recognition technology. Journal of Dialectics of Nature Communication, 45(2), 82-90. https://doi.org/10.15994/j.1000-0763.2023.02.011

[19] Yang, Y. (2026). Risk regulation of emotion recognition systems: The EU model and its implications. Journal of Huazhong University of Science and Technology (Social Sciences Edition), 40(2), 40-53. https://doi.org/10.19648/j.cnki.jhustss1980.2026.02.04

[20] Verma, A., & Singh, L. K. (2025). Hybrid feature extraction using LBP and HOG for facial emotion recognition.

[21] Choudhury, A., & Bhowmik, T. (2025). Speech emotion recognition using deep learning techniques.

[22] Tao, F., & Liu, G. (2018). Advanced LSTM: A study about better time dependency modeling in emotion recognition. In ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 2906-2910). IEEE.

[23] Wang, H., Hua, Y., Zhang, L., & Hakim, M. N. (2024). Speech emotion recognition of Guanzhong dialect based on CNN and BiLSTM. In 2024 9th International Conference on Intelligent Computing and Signal Processing (ICSP) (pp. 812–815). IEEE. https://doi.org/10.1109/ICSP62122.2024.10743655

[24] Salur, M. U., & Kaymaz, M. (2025). Facial emotion recognition using vision transformer enhanced with attention mechanisms. In 2025 9th International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE. https://doi.org/10.1109/IDAP68205.2025.11222259.

[25] Buvana, J., Rajalakshmi, S., Yeseswini, S., Vandana, E., & Lathika, S. P. (2025). Wearable-assisted multimodal emotion recognition for stress and neurological disorder monitoring. In 2025 IEEE 6th Global Conference for Advancement in Technology (GCAT). IEEE. https://doi.org/10.1109/GCAT66372.2025.11368349

[26] Sutoyo, R., et al. (2023). The impact of emotion recognition models towards believability factor of chatbots.