Short Note

Embodied AI: Multimodal Integration of Facial Expressions and Biometric Signals

Thabo Mosala *

Posted Date: 3 September 2025

doi: 10.20944/preprints202509.0270.v1

Keywords: embodied AI; affective computing; multimodal fusion; emotion recognition; biometric signals; facial expression analysis; adaptive coaching

 doi:10.20944/preprints202509.0270.v1

Short Note

Embodied AI: Multimodal Integration of Facial Expressions and Biometric Signals

Thabo Mosala

Wits Business School, Parktown, Johannesburg, South Africa; thabo.mosala@wits.ac.za

Abstract

Artificial Intelligence (AI) systems increasingly support human development in domains such as coaching, education, and healthcare. Yet most remain disembodied, relying solely on text or speech while neglecting non-verbal cues that are central to human communication. This research advances the science of embodied AI by developing computational models that integrate facial expressions and biometric signals (heart rate, HRV, temperature, electrodermal activity) for robust, real-time affect recognition. Building on embodied cognition, polyvagal theory, multimodal machine learning, and affective computing, the study compares early, late, and hybrid fusion architectures for synchronizing heterogeneous data streams. A mixed evaluation design combines benchmarking against multimodal datasets with experimental trials in adaptive coaching contexts. The expected contribution is twofold: (1) scientific, novel multimodal fusion architectures and comparative insights into fusion trade-offs; and (2) applied, an embodied AI coaching prototype and ethical guidelines for biometric data use. This work bridges gaps in affective computing and paves the way for emotionally intelligent, context-aware AI systems.

Keywords: embodied AI; affective computing; multimodal fusion; emotion recognition; biometric signals; facial expression analysis; adaptive coaching

1. Introduction

Artificial Intelligence (AI) has advanced in natural language processing, computer vision, and speech recognition (Bian et al., 2023). Yet, most human-facing AI systems—such as those in coaching, education, or healthcare—remain disembodied, relying primarily on text or voice (McKee et al., 2023; Modi & Devaraj, 2022; Peng et al., 2024). Human communication, however, is inherently multimodal, combining language, facial micro-expressions, and involuntary physiological responses (Liu et al., 2023; Tiwari & Falk, 2019).

To achieve empathetic and adaptive interaction, AI must interpret these non-verbal signals (Spitale et al., 2024). This paper proposes the development of embodied AI systems that integrate facial expression recognition with biometric data (heart rate, heart rate variability, temperature, electrodermal activity) to enhance affect recognition in real time (Gao et al., 2024; Suganya et al., 2024). Such integration strengthens AI's emotional intelligence, allowing adaptive coaching systems to tailor interventions dynamically to a user's emotional and physiological state (Liu, 2024; Awan et al., 2022).

While affective computing has pioneered unimodal approaches (Picard, 1997; Ekman & Friesen, 1978), challenges remain in multimodal fusion, synchronization, and contextual interpretation (Wang et al., 2022; Yang et al., 2024). This research advances multimodal AI by developing and testing fusion architectures capable of operating in dynamic, real-world environments such as adaptive coaching (Narimisaei et al., 2024; Zhu et al., 2025).

2. Problem Statement and Research Questions

 doi:10.20944/preprints202509.0270.v1

Despite rapid advances in artificial intelligence, most affect recognition systems remain **unimodal**, relying on either text, voice, or facial expressions alone. This creates an **empathy gap**, as human communication is inherently multimodal, blending facial micro-expressions, body language, and physiological signals (Mehrabian, 1971; Ekman & Friesen, 1978; Liu et al., 2023). Unimodal systems therefore struggle with **accuracy** and **context sensitivity**, often misinterpreting emotional states in real-world conditions (Mathur et al., 2023; Poria et al., 2017; Afzal et al., 2024).

A second limitation is **synchronization**: biometric and visual signals operate at different temporal resolutions, making it difficult to align and fuse them effectively (Wang et al., 2022; Shakhovska et al., 2024; Zhu et al., 2025). Without robust temporal integration, critical affective cues may be lost. Finally, most models remain **context-agnostic**, capable of detecting emotions but not dynamically adjusting their responses in ways that mirror human social intelligence (Mohamed et al., 2024; Robb et al., 2023). This undermines user trust and engagement in sensitive domains such as coaching, therapy, and education (Nyamathi et al., 2024; Spitale et al., 2024).

Addressing these gaps requires developing **multimodal frameworks** that integrate facial and biometric signals into adaptive feedback loops. Such systems would not only improve recognition accuracy but also enable **embodied AI** capable of more empathetic, context-aware interactions (Picard, 1997; Baltrušaitis et al., 2019; Hegde & Jayalath, 2025).

Accordingly, this study poses three guiding research questions:

RQ1: How can multimodal deep learning architectures effectively fuse asynchronous data streams (facial micro-expressions, heart rate, HRV, temperature) for robust real-time affect recognition?

RQ2: What are the computational trade-offs between early, late, and hybrid fusion approaches in multimodal affect detection, particularly under real-time constraints?

RQ3: How can multimodal affect recognition be embedded into adaptive AI systems that dynamically modify responses in interactive coaching environments?

3. Literature Review

Research in **affective computing** has long emphasized the importance of enabling machines to detect and respond to human emotions (Picard, 1997). Subsequent surveys have highlighted its applications in education, healthcare, and human–AI interaction but note persistent challenges in moving beyond unimodal designs (Calvo & D'Mello, 2010; Poria et al., 2017; Afzal et al., 2024; Vistorte et al., 2024; Mathur et al., 2023; Hegde & Jayalath, 2025). **Computer vision approaches** grounded in Ekman and Friesen's (1978) Facial Action Coding System have shown that micro-expressions reveal hidden affective states, yet models remain vulnerable to occlusion, lighting, and cultural variability (Bian et al., 2023; Huang et al., 2023; Janhonen, 2023; Tellamekala et al., 2025). Similarly, **physiological computing** has linked biometric signals such as heart rate variability and electrodermal activity to stress, arousal, and regulation (Porges, 2011; Beatton et al., 2024; Mattern et al., 2023; Pessanha & Salah, 2021). However, these signals are often noisy and context-dependent, limiting their reliability when used in isolation (Bello et al., 2023; Wang & Wang, 2025).

To address these shortcomings, **multimodal fusion** approaches integrate visual, auditory, and physiological signals. Landmark surveys (Baltrušaitis et al., 2019; Koromilas & Γιαννακόπουλος, 2021; Lai et al., 2023; Li et al., 2025) show that fusion improves robustness and accuracy compared to unimodal systems. Recent innovations include cross-modal attention (Das et al., 2024), latent distribution calibration (Tellamekala et al., 2023), and interpretable fusion frameworks (Mansouri-Benssassi & Ye, 2021; Zhi et al., 2024). Empirical work has further explored multimodal sentiment analysis (Pan & Liu, 2024), audio–visual emotion recognition (Schoneveld et al., 2021), and real-time estimation using behavioral and neurophysiological signals (Herbuela & Nagai, 2025; Mordacq et al., 2024). Yet, major technical barriers remain: asynchronous signals complicate alignment (Wang et al., 2022; Shakhovska et al., 2024; Zhu et al., 2025), scalability challenges hinder deployment in dynamic settings (Gupta et al., 2024; Bose et al., 2023), and ethical concerns persist regarding biometric and facial data (Barker et al., 2025; Afroogh et al., 2024; Chavan et al., 2025; Lin, 2024).

 doi:10.20944/preprints202509.0270.v1

Despite these advances, few studies have delivered **computationally robust, real-time frameworks** that integrate visual and biometric cues within adaptive, interactive systems. Current models often detect affective states but lack the capacity to dynamically adjust responses in ways that mirror human social intelligence (Mohamed et al., 2024; Robb et al., 2023; Niebuhr & Valls-Ratés, 2024). This gap underscores the need for a framework that fuses **facial expressions and biometric signals** into **adaptive feedback loops** for embodied AI. Such a system would bridge unimodal limitations, overcome synchronization challenges, and deliver more empathetic, context-aware AI agents—particularly in coaching, education, and therapeutic settings (Alazraki et al., 2021; Dol et al., 2023; Nyamathi et al., 2024; Kok et al., 2024; Hao et al., 2024).

4. Theoretical and Computational Framework

This study draws on four complementary frameworks to guide the development of multimodal affect recognition and adaptive coaching systems:

Embodied Cognition Theory

Embodied cognition posits that cognition and emotion are inseparable from bodily states, meaning that affect must be understood through physiological and behavioral signals. Lakoff and Johnson (1999) introduced embodiment as central to cognition, while Wilson (2002) outlined six perspectives that shaped the field. Barsalou (2008) further advanced the concept through grounded cognition, showing that abstract thought is rooted in bodily simulation. These perspectives support **RQ1**, justifying why multimodal AI systems should fuse facial and biometric data to approximate human affective understanding (Klippel et al., 2021; Hauke et al., 2024; Liu et al., 2023).

Polyvagal Theory

Polyvagal theory links autonomic physiology, particularly heart rate variability (HRV), to readiness for social engagement and stress regulation (Porges, 2011). This aligns with **RQ2**, providing a foundation for interpreting biometric signals. Gross (1998) emphasized emotion regulation as a process shaped by both physiology and cognition, while Cacioppo, Tassinary, and Berntson (2007) established psychophysiology as a scientific basis for understanding emotional states. These works collectively justify the inclusion of HRV and electrodermal activity as inputs to multimodal fusion (Beatton et al., 2024; Puglisi et al., 2023; Lee et al., 2023; Herbuela & Nagai, 2025).

Multimodal Machine Learning Frameworks

Multimodal machine learning provides computational strategies for integrating heterogeneous signals. Early surveys emphasized decision and feature-level fusion (Atrey et al., 2010), while Ngiam et al. (2011) introduced deep multimodal learning, paving the way for current neural architectures. Baltrušaitis et al. (2019) synthesized advances into a widely cited taxonomy. These foundations inform **RQ1 and RQ2**, guiding evaluation of early, late, and hybrid fusion strategies. Recent developments—such as calibrated latent distribution fusion (Tellamekala et al., 2023), meta-fusion frameworks (Liang et al., 2025), and interpretable fusion models (Zhi et al., 2024; Zhu et al., 2025)—extend this foundation toward robust, explainable real-time integration.

Affective Computing Paradigm

Picard's (1997) seminal work established affective computing, later extended by Schröder and Cowie (2005), who highlighted design challenges for emotion-oriented systems. Calvo and D'Mello (2010) reviewed affect detection methods, while D'Mello and Kory (2015) provided a meta-analysis of multimodal affect recognition in learning environments. These foundational works support **RQ3**, which explores embedding affect recognition into adaptive feedback loops. Recent studies advance these ideas by applying affective computing to real-world domains, including wellbeing (Spitale et al., 2024), learning (Vistorte et al., 2024), and emotional support (Hegde & Jayalath, 2025; Mohamed et al., 2024).

Collectively, these frameworks argue that embodied AI requires integration of bodily, physiological, and affective signals, computationally modeled through multimodal machine learning and affective computing to achieve emotionally intelligent, adaptive interactions.

 doi:10.20944/preprints202509.0270.v1

5. Proposed Methodology

This research employs a mixed-methods, quasi-experimental design to develop and evaluate multimodal fusion architectures for affect recognition, with a specific focus on adaptive coaching applications. The methodology combines computational modeling, empirical testing, and comparative evaluation to address the research questions.

Model Development

Three deep learning fusion architectures will be developed and benchmarked:

- Early Fusion: Raw biometric and visual signals will be combined prior to feature extraction, following approaches tested in multimodal affect detection (Tellamekala et al., 2023; Wang et al., 2022).
- Late Fusion: Independent unimodal models will be trained and merged at the decision stage, leveraging ensemble and evidential methods (El-Din et al., 2023; Liang et al., 2025).
- Hybrid Fusion: A shared feature space with cross-attention mechanisms will be implemented to dynamically integrate complementary cues, drawing on advances in interpretable fusion frameworks (Mansouri-Benssassi & Ye, 2021; Shakhovska et al., 2024; Zhao et al., 2021).

Each architecture will be embedded in a prototype Embodied AI Coach/therapy to test real-time adaptability. The framework builds on prior work in multimodal sentiment classification (Suganya et al., 2024), audio-visual emotion recognition (Schoneveld et al., 2021), and bio-inspired computational integration (Mansouri-Benssassi & Ye, 2021).

Participants and Data Collection

A sample of **40–60 participants** will engage in structured coaching sessions using both unimodal (baseline) and multimodal (embodied) AI systems, consistent with prior experimental affective computing designs (Nyamathi et al., 2024; Robb et al., 2023).

Data sources include:

- Facial micro-expressions via video capture (Ekman & Friesen, 1978; Huang et al., 2023).
- Biometric signals including heart rate variability, galvanic skin response, and skin temperature (Pessanha & Salah, 2021; Beatton et al., 2024; Mattern et al., 2023).
- Self-report surveys to assess empathy, trust, and satisfaction (Fang et al., 2023; Harris et al., 2023).
- System logs capturing adaptive responses, latency, and feedback timing (Shore et al., 2023).

This multimodal dataset design aligns with large-scale emotion corpora such as K-EmoCon (Park et al., 2020) and mixed emotion datasets (Yang et al., 2024).

Data Analysis

The evaluation integrates quantitative and qualitative methods:

- Quantitative Analysis: Statistical comparisons (accuracy, precision, recall, F1-scores, ANOVA) will test the performance of early, late, and hybrid fusion approaches under real-time constraints (Wu et al., 2023; Hassan et al., 2025).
- Qualitative Analysis: Thematic coding of user reflections will assess perceived empathy, effectiveness, and trust (Niebuhr & Valls-Ratés, 2024; Rossing et al., 2024).
- Computational Trade-offs: The study will measure efficiency, interpretability, and scalability of different fusion strategies in dynamic environments (Bian et al., 2023; Bose et al., 2023; Liao et al., 2025).

This design ensures that the methodology addresses both scientific objectives—evaluating multimodal integration architectures—and applied outcomes—demonstrating their potential in

 doi:10.20944/preprints202509.0270.v1

embodied AI coaching systems, while remaining ethically grounded in biometric and facial data use (Afroogh et al., 2024; Lin, 2024).

Develop and compare three fusion architectures:

- Early Fusion – raw biometric + visual inputs combined before feature extraction (Tellamekala et al., 2023; Wang et al., 2022).
- Late Fusion – unimodal outputs merged at decision level (Liang et al., 2025; El-Din et al., 2023).
- Hybrid Fusion – feature-level concatenation with cross-attention for adaptive weighting (Mansouri-Benssassi & Ye, 2021; Song et al., 2024; Wu et al., 2023).

Evaluation

- Data: facial video (micro-expressions) + physiological signals (HRV, GSR, temperature).
- Benchmarks: accuracy, F1, latency, interpretability across datasets (DEAP, MAHNOB-HCI).
- Metrics: precision, recall, computational efficiency, robustness across populations (Gupta et al., 2024; Hassan et al., 2025).

6. Expected Contribution

This research offers contributions on both the **scientific** and **applied** fronts, advancing the design of embodied AI for affective computing and adaptive coaching.

Scientific Contributions

- Development of **novel fusion architectures** (early, late, and hybrid) to integrate facial expressions with biometric signals for robust multimodal affect recognition.
- **Comparative evaluation** of fusion strategies under real-time constraints, addressing performance, accuracy, and scalability challenges in dynamic environments.
- Establishment of a **computational framework** linking multimodal affect recognition to adaptive decision-making, thereby deepening the scientific understanding of embodied AI.

Applied Contributions

- Design and testing of a **prototype Embodied AI Coach** capable of delivering real-time adaptive feedback informed by users' affective and physiological states.
- Practical **insights for deploying multimodal AI** in education, coaching, and therapy, highlighting opportunities for more personalized and empathetic interventions.
- Development of **ethical guidelines** for the collection and use of facial and biometric data, supporting responsible innovation and safeguarding user trust.

Together, these contributions extend the **science of multimodal affect recognition** while providing a pathway to **practical, ethically grounded applications** in coaching and human–AI interaction.

7. Limitations and Future Research

Dataset Dependence.

The study relies on existing multimodal datasets (e.g., DEAP, MAHNOB-HCI, K-EmoCon), which, while widely used, present constraints in terms of sample diversity, ecological validity, and cultural variability (Pan & Liu, 2024; Yang et al., 2024). Many datasets are lab-controlled and lack the contextual noise of real-world environments, limiting generalization to natural coaching or therapy contexts. Future research should extend benchmarking to more **ecologically valid datasets** and collect new multimodal corpora in real-world adaptive interactions.

Fusion Complexity and Synchronization.

 doi:10.20944/preprints202509.0270.v1

Although early, late, and hybrid fusion strategies provide a framework for integration, asynchronous sampling rates and noise across modalities complicate synchronization (Wang et al., 2022; Shakhovska et al., 2024; Zhu et al., 2025). Deep models risk overfitting to dataset-specific noise and may lack robustness when deployed in real-time applications. Future research should investigate **cross-attention mechanisms, adaptive alignment algorithms, and explainable AI methods** to ensure stability and interpretability of multimodal fusion models.

Computational Efficiency and Scalability.

Real-time multimodal affect recognition requires significant computational resources, creating trade-offs between accuracy, latency, and scalability (Bose et al., 2023; Gupta et al., 2024). Prototype models may perform well in experimental settings but struggle under deployment constraints such as low-power devices or bandwidth-limited contexts. Future work should prioritize **lightweight architectures, edge-AI implementations, and energy-efficient models** to enhance scalability in diverse environments.

Ethical and Interpretability Challenges.

While this study emphasizes technical performance, integrating sensitive biometric data introduces **ethical concerns** regarding privacy, security, and potential misuse (Afroogh et al., 2024; Chavan et al., 2025). Moreover, deep fusion models often act as "black boxes," limiting interpretability for end-users and practitioners. Future research should explore **explainable multimodal AI**, ensuring that both researchers and users understand how emotional inferences are made.

Future Research Directions.

Moving forward, research should expand in four directions:

1. **Cross-domain generalization** — testing architectures in varied applied domains such as education, healthcare, and workplace coaching.
2. **Longitudinal evaluation** — measuring how multimodal models adapt to user changes over time rather than single-session trials.
3. **Neuro-inspired integration** — leveraging bio-inspired computational models (Mansouri-Benssassi & Ye, 2021) to improve affective state modeling.
4. **Ethical frameworks** — developing governance standards for multimodal AI systems to ensure responsible deployment in sensitive human-centered contexts.

References

1. Afzal, S., Khan, H. A., Piran, M. J., & Lee, J. W. (2024). A comprehensive survey on affective computing: Challenges, trends, applications, and future directions. IEEE Access, 12, 96150. https://doi.org/10.1109/ACCESS.2024.3422480
2. Alazraki, L., Ghachem, A., Polydorou, N., Khosmood, F., & Edalat, A. (2021). An empathetic AI coach for self-attachment therapy. 2021 IEEE 3rd International Conference on Cognitive Machine Intelligence (CogMI), 78–85. https://doi.org/10.1109/COGMI52975.2021.00019
3. Ali, K., & Hughes, C. E. (2023). A unified transformer-based network for multimodal emotion recognition. TechRxiv. https://doi.org/10.36227/techrxiv.23916123.v1
4. Andrews, J. T. A., Zhao, D., Thong, W., Modas, A., Papakyriakopoulos, O., & Xiang, A. (2023). Ethical considerations for responsible data curation. arXiv. https://doi.org/10.48550/arXiv.2302.03629
5. Arsalan, A., Anwar, S. M., & Majid, M. (2022). Human stress assessment: A comprehensive review of methods using wearable sensors and non-wearable techniques. arXiv. https://doi.org/10.48550/arxiv.2202.03033
6. Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. ACM Computing Surveys, 42(2), 1–37. https://doi.org/10.1145/1670679.1670680
7. Awan, A. W., Usman, S. M., Khalid, S., Anwar, A., Alroobaea, R., Hussain, S., Almotiri, J., Ullah, S. S., & Akram, M. U. (2022). An ensemble learning method for emotion charting using multimodal physiological signals. Sensors, 22(23), 9480. https://doi.org/10.3390/s22239480

 doi:10.20944/preprints202509.0270.v1

8. Barker, D., Tippireddy, M. K. R., Farhan, A., & Ahmed, B. (2025). Ethical considerations in emotion recognition research. Psychology International, 7(2), 43. https://doi.org/10.3390/psycholint7020043
9. Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639
10. Bassi, G., Giuliano, C., Perinelli, A., Forti, S., Gabrielli, S., & Salcuni, S. (2021). A virtual coach (Motibot) for supporting healthy coping strategies among adults with diabetes: Proof-of-concept study. JMIR Human Factors, 9(1). https://doi.org/10.2196/32211
11. Bello, H., Marin, L. A. S., Suh, S., Zhou, B., & Lukowicz, P. (2023). InMyFace: Inertial and mechanomyography-based sensor fusion for wearable facial activity recognition. Information Fusion, 99, 101886. https://doi.org/10.1016/j.inffus.2023.101886
12. Bian, Y., Küster, D., Liu, H., & Krumhuber, E. G. (2023). Understanding naturalistic facial expressions with deep learning and multimodal large language models. Sensors, 24(1), 126. https://doi.org/10.3390/s24010126
13. Bose, D., Hebbar, R., Somandepalli, K., & Narayanan, S. (2023). Contextually-rich human affect perception using multimodal scene information. ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095728
14. Bota, P., Wang, C., Fred, A., & Silva, H. (2020). Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet? Sensors, 20(17), 4723. https://doi.org/10.3390/s20174723
15. Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). Handbook of psychophysiology (3rd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511546396
16. Calvo, R. A., & D'Mello, S. K. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing, 1(1), 18–37. https://doi.org/10.1109/T-AFFC.2010.1
17. Chavan, V., Cenaj, A., Shen, S., Bar, A., Binwani, S., Del Becaro, T., Funk, M., Greschner, L., Hung, R., Klein, S., Kleiner, R., Krause, S., Olbrych, S., Parmar, V., Sarafraz, J., Soroko, D., Don, D. W., Zhou, C., Vu, H. T. D., … Fresquet, X. (2025). Feeling machines: Ethics, culture, and the rise of emotional AI. arXiv. https://doi.org/10.48550/arXiv.2506.12437
18. Chen, J., Wang, X., Huang, C., Hu, X., Shen, X., & Zhang, D. (2023). A large finer-grained affective computing EEG dataset. Scientific Data, 10(1). https://doi.org/10.1038/s41597-023-02650-w
19. Das, A., Sarma, M. S., Hoque, M. M., Siddique, N., & Dewan, M. A. A. (2024). AVaTER: Fusing audio, visual, and textual modalities using cross-modal attention for emotion recognition. Sensors, 24(18), 5862. https://doi.org/10.3390/s24185862
20. D'Mello, S., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. ACM Computing Surveys, 47(3), 1–36. https://doi.org/10.1145/2682899
21. Dol, A., van Strien, T., Velthuijsen, H., van Gemert-Pijnen, J. E. W. C., & Bode, C. (2023). Preferences for coaching strategies in a personalized virtual coach for emotional eaters: An explorative study. Frontiers in Psychology, 14, 1260229. https://doi.org/10.3389/fpsyg.2023.1260229
22. Ekman, P., & Friesen, W. V. (1978). Facial Action Coding System: A technique for the measurement of facial movement. Consulting Psychologists Press.
23. Fang, Y., Rong, R. M., & Huang, J. (2021). Hierarchical fusion of visual and physiological signals for emotion recognition. Multidimensional Systems and Signal Processing, 32(4), 1103–1124. https://doi.org/10.1007/s11045-021-00774-z
24. Gao, R., Liu, X., Xing, B., Yu, Z., Schuller, B. W., & Kälviäinen, H. (2024). Identity-free artificial emotional intelligence via micro-gesture understanding. arXiv. https://doi.org/10.48550/arXiv.2405.13206
25. Goleman, D. (1995). Emotional intelligence: Why it can matter more than IQ. Bantam Books.
26. Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. Review of General Psychology, 2(3), 271–299. https://doi.org/10.1037/1089-2680.2.3.271
27. Guntz, T. (2020). Estimating expertise from eye gaze and emotions. HAL. https://theses.hal.science/tel-03026375
28. Gupta, N., Priya, R. V., & Verma, C. K. (2024). ERFN: Leveraging context for enhanced emotion detection. International Journal of Advanced Computer Science and Applications, 15(6). https://doi.org/10.14569/IJACSA.2024.0150663

 doi:10.20944/preprints202509.0270.v1

29. Habibi, R., Pfau, J., Holmes, J., & El-Nasr, M. S. (2023). Empathetic AI for empowering resilience in games. arXiv. https://doi.org/10.48550/arXiv.2302.09070
30. Hao, F., Zhang, H., Wang, B., Liao, L., Liu, Q., & Cambria, E. (2024). EmpathyEar: An open-source avatar multimodal empathetic chatbot. arXiv. https://doi.org/10.48550/arXiv.2406.15177
31. Hassan, A., Ahmad, S. G., Iqbal, T., Munir, E. U., Ayyub, K., & Ramzan, N. (2025). Enhanced model for gestational diabetes mellitus prediction using a fusion technique of multiple algorithms with explainability. International Journal of Computational Intelligence Systems, 18(1). https://doi.org/10.1007/s44196-025-00760-4
32. Hauke, G., Lohr-Berger, C., & Shafir, T. (2024). Emotional activation in a cognitive behavioral setting: Extending the tradition with embodiment. Frontiers in Psychology, 15, 1409373. https://doi.org/10.3389/fpsyg.2024.1409373
33. Hegde, K., & Jayalath, H. (2025). Emotions in the loop: A survey of affective computing for emotional support. arXiv. https://doi.org/10.48550/arXiv.2505.01542
34. Herbuela, V. R. D. M., & Nagai, Y. (2025). Realtime multimodal emotion estimation using behavioral and neurophysiological data. arXiv. https://doi.org/10.48550/arXiv.2508.09402
35. Huang, Z., Chiang, C., Chen, J., Chen, Y.-C., Chung, H.-L., Cai, Y., & Hsu, H.-C. (2023). A study on computer vision for facial emotion recognition. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-35446-4
36. Islam, R., & Bae, S. W. (2024). Revolutionizing mental health support: An innovative affective mobile framework for dynamic, proactive, and context-adaptive conversational agents. arXiv. https://doi.org/10.48550/arXiv.2406.15942
37. Janhonen, J. (2023). Socialisation approach to AI value acquisition: Enabling flexible ethical navigation with built-in receptiveness to social influence. AI and Ethics. https://doi.org/10.1007/s43681-023-00372-8
38. Lakoff, G., & Johnson, M. (1999). Philosophy in the flesh: The embodied mind and its challenge to Western thought. Basic Books.
39. Lee, S., & Kim, J.-H. (2023). Video multimodal emotion recognition system for real world applications. arXiv. https://doi.org/10.48550/arXiv.2308.14320
40. Li, Y., Sun, Q., Schlicher, M., Lim, Y. W., & Schuller, B. W. (2025). Artificial emotion: A survey of theories and debates on realising emotion in artificial intelligence. arXiv. https://doi.org/10.48550/arXiv.2508.10286
41. Liao, Y., Gao, Y., Wang, F., Zhang, L., Xu, Z., & Wu, Y. (2025). Emotion recognition with multiple physiological parameters based on ensemble learning. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-96616-0
42. Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. Sluyter (Eds.), Emotional development and emotional intelligence: Educational implications (pp. 3–31). Basic Books.
43. McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. iScience, 26(8), 107256. https://doi.org/10.1016/j.isci.2023.107256
44. Mehrabian, A. (1971). Silent messages. Wadsworth.
45. Modi, K., & Devaraj, L. (2022). Advancements in biometric technology with artificial intelligence. arXiv. https://doi.org/10.48550/arxiv.2212.13187
46. Narimisaei, J., Naeim, M., Imannezhad, S., Samian, P., & Sobhani, M. (2024). Exploring emotional intelligence in AI systems: A comprehensive analysis of emotion recognition and response mechanisms. Annals of Medicine and Surgery, 86(8), 4657. https://doi.org/10.1097/MS9.0000000000002315
47. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (pp. 689–696). Omnipress.
48. Nyamathi, A., Dutt, N., Lee, J., Rahmani, A. M., Rasouli, M., Krogh, D., Krogh, E., Sultzer, D. L., Rashid, H., Liaqat, H., Jawad, R., Azhar, F., Ali, A., Qamar, B., Bhatti, T. Y., Khay, C., Ludlow, J., Gibbs, L., Rousseau, J., … Brunswicker, S. (2024). Establishing the foundations of emotional intelligence in care companion robots to mitigate agitation among high-risk patients with dementia: Protocol for an empathetic patient-robot interaction study. JMIR Research Protocols, 13, e55761. https://doi.org/10.2196/55761
49. Pan, L., & Liu, W. (2024). Adaptive language-interacted hyper-modality representation for multimodal sentiment analysis. International Journal of Advanced Computer Science and Applications, 15(7). https://doi.org/10.14569/IJACSA.2024.0150746
50. Picard, R. W. (1997). Affective computing. MIT Press. https://doi.org/10.7551/mitpress/1145.001.0001
51. Porges, S. W. (2011). The polyvagal theory: Neurophysiological foundations of emotions, attachment, communication, and self-regulation. W. W. Norton & Company.

 doi:10.20944/preprints202509.0270.v1

52. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003
53. Schröder, M., & Cowie, R. (2005). Issues in emotion-oriented computing. In R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. F. Papageorgiou, S. Kollias, & S. O. O'Donnell (Eds.), Emotion-Oriented Systems (pp. 3–18). Springer. https://doi.org/10.1007/1-4020-2760-9_1
54. Spitale, M., Winkle, K., Barakova, E., & Güneş, H. (2024). Guest editorial: Special issue on embodied agents for wellbeing. International Journal of Social Robotics, 16(5), 833–838. https://doi.org/10.1007/s12369-024-01150-0
55. Suganya, R., Narmatha, M., & Kumar, S. V. (2024). An emotionally intelligent system for multimodal sentiment classification. Indian Journal of Science and Technology, 17(42), 4386–4396. https://doi.org/10.17485/ijst/v17i42.2349
56. Tiwari, A., & Falk, T. H. (2019). Fusion of motif- and spectrum-related features for improved EEG-based emotion recognition. Computational Intelligence and Neuroscience, 2019, 1–15. https://doi.org/10.1155/2019/3076324
57. Vistorte, A. O. R., Deroncele-Acosta, Á., Ayala, J. L. M., Barrasa, Á., López-Granero, C., & Martí-González, M. (2024). Integrating artificial intelligence to assess emotions in learning environments: A systematic literature review. Frontiers in Psychology, 15, 1387089. https://doi.org/10.3389/fpsyg.2024.1387089
58. Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion, 19, 29–55. https://doi.org/10.1016/j.inffus.2022.03.009
59. Yang, P., Liu, N., Liu, X., Shu, Y., Ji, W., Ren, Z., Sheng, J., Yu, M., Yi, R., Zhang, D., & Liu, Y. (2024). A multimodal dataset for mixed emotion recognition. Scientific Data, 11(1). https://doi.org/10.1038/s41597-024-03676-4
60. Zhang, L., Qian, Y., Arandjelović, O., & Zhu, A. (2023). Multimodal latent emotion recognition from micro-expression and physiological signals. arXiv. https://doi.org/10.48550/arxiv.2308.12156
61. Zhao, S., Jia, G., Yang, J., Ding, G., & Keutzer, K. (2021). Emotion recognition from multiple modalities: Fundamentals and methodologies. IEEE Signal Processing Magazine, 38(6), 59–73. https://doi.org/10.1109/MSP.2021.3106895
62. Zhi, H., Hong, H., Cai, X., Li, L., Ren, Z., Xiao, M., Jiang, H., & Wang, X. (2024). Skew-pair fusion theory: An interpretable multimodal fusion framework. Research Square. https://doi.org/10.21203/rs.3.rs-5208094/v1
63. Zhu, Y., Han, L., Jiang, G., Zhou, P., & Wang, Y. (2025). Hierarchical MoE: Continuous multimodal emotion recognition with incomplete and asynchronous inputs. arXiv. https://doi.org/10.48550/arXiv.2508.02133