Future Technology
Open Access Journal
ISSN 2832-0379

February 2026| Volume 05 | Issue 01 | Pages 65-71

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.1.7

Future Publishing LLC

Article

Multimodal emotion recognition-driven personalized digital therapeutics for anxiety management

Lusha Zhu, Jinho Yim*

Department of Smart Experience Design, Kookmin University, Seoul 01706, Republic of Korea

ARTICLE INFO

Article history:
Received 20 August 2025
Received in revised form
01 October 2025
Accepted 17 October 2025

Keywords:
Multimodal emotion recognition,
Digital therapeutics, Anxiety,
Reinforcement learning,
Personalized intervention

*Corresponding author
Email address:
hci.yim@kookmin.ac.kr

DOI: 10.55670/fpll.futech.5.1.7

ABSTRACT

Anxiety disorders are among the most widespread mental health challenges, yet conventional treatments face barriers of accessibility, cost, and reliance on subjective measures. Digital therapeutics offer scalable solutions, but current systems lack real-time emotion monitoring and adaptive personalization. To address this gap, this study proposes a multimodal emotion recognition-driven framework for personalized anxiety management. The framework fuses electroencephalography, heart rate variability, facial expression, and speech features via cross-modal attention, and employs a reinforcement learning-based decision engine to dynamically select interventions such as breathing exercises, mindfulness, or cognitive reframing. Adaptive feedback further tailors interventions to user responses. Experiments on DEAP and WESAD datasets showed superior performance over unimodal and traditional fusion baselines, with accuracies of 86.2% and 84.7% and AUROCs of 0.91 and 0.89. Anxiety reduction analysis demonstrated up to 24% improvement in State-Trait Anxiety Inventory scores. The study advances affective computing by linking multimodal sensing with adaptive therapeutic design, and offers a foundation for scalable, interpretable, and clinically relevant digital mental health interventions.

1. Introduction

Anxiety disorders are among the most prevalent mental health conditions, affecting nearly one in five individuals worldwide and imposing substantial social and economic burdens [1]. Conventional treatments such as pharmacological therapy and cognitive-behavioral interventions are clinically effective but limited by accessibility, high cost, and stigma [2]. In this context, digital therapeutics (DTx) have emerged as scalable, data-driven alternatives that leverage mobile and intelligent systems to deliver evidence-based care [3]. Meanwhile, recent progress in artificial intelligence (AI)-driven emotion recognition offers new opportunities to enhance DTx through objective, real-time monitoring of emotional states and adaptive intervention delivery [4]. Despite these advances, current anxiety-focused DTx largely depend on self-report data and static engagement metrics that fail to capture rapid emotional fluctuations. Multimodal emotion recognition (MER) research combining physiological (EEG, HRV), facial, and vocal cues has achieved higher accuracy than unimodal approaches [5], yet most studies remain laboratory-bound and emphasize recognition accuracy rather than therapeutic adaptation. This gap underscores the need for systems that link multimodal affect sensing with intelligent control mechanisms capable of autonomously selecting and adjusting therapeutic actions [6]. To address this gap, this study reconceptualizes emotion recognition not as an endpoint but as a dynamic input for adaptive digital therapeutics. The proposed framework integrates a cross-modal attention mechanism for robust affect fusion with a reinforcement-learning engine that learns optimal intervention policies from user feedback. This synergy between perception and decision-making constitutes the core methodological novelty and distinguishes the work from prior static DTx architectures. From a technological standpoint, the framework advances AI-driven health systems through a unified architecture that combines cross-modal attention, actor-critic reinforcement learning, and adaptive feedback loops. These modules enable continuous personalization and can be efficiently deployed on mobile or wearable platforms for real-time, privacy-preserving inference. The design demonstrates how engineering innovations in multimodal fusion and policy optimization can yield clinically meaningful outcomes without sacrificing computational scalability. The objectives of this research are threefold: (1) to develop a multimodal emotion-recognition model that fuses heterogeneous physiological and behavioral

data for reliable real-time affect inference; (2) to design a reinforcement-learning-based decision engine that dynamically selects personalized interventions; and (3) to evaluate the proposed framework through quantitative benchmarks and anxiety-reduction analyses demonstrating both accuracy and therapeutic efficacy. Methodologically, multimodal data, including EEG, HRV, facial expression, and speech, are pre-processed for feature extraction and fused via cross-modal attention. The resulting embeddings inform a reinforcement-learning agent that adaptively recommends therapeutic activities such as breathing exercises, mindfulness prompts, or cognitive reframing. Experimental validation includes comparisons with unimodal and fusion baselines, ablation tests, and statistical significance analyses to verify robustness and interpretability. In summary, this study contributes to the technological advancement of affective computing by transforming multimodal emotion recognition into an adaptive control signal for personalized digital therapeutics. It provides a reproducible AI architecture that unites multimodal sensing, real-time learning, and human-centered feedback, offering a scalable pathway toward next-generation, intelligent, and ethically deployable digital interventions for anxiety management.

2. Related works

Research on anxiety-oriented digital therapeutics intersects three major domains: multimodal emotion recognition, digital interventions for anxiety management, and personalized adaptive strategies. Each domain has generated significant progress, yet notable limitations remain, highlighting the need for integrated solutions.

2.1 Multimodal emotion recognition

Recent years have witnessed significant advances in MER, which integrates physiological signals, visual cues, and vocal features to achieve superior accuracy compared with unimodal methods. Physiological modalities such as EEG and HRV offer objective measures of affective states, while facial expression analysis and speech prosody provide complementary behavioral cues [7]. Deep learning models, including convolutional and recurrent architectures, have been widely applied, with fusion strategies ranging from early concatenation to attention-based cross-modal integration. Although these approaches demonstrate improved recognition rates on benchmark datasets, challenges persist. The majority of studies rely on controlled laboratory environments, resulting in reduced robustness in naturalistic settings. Moreover, most work prioritizes classification accuracy over explainability, limiting clinical interpretability [8]. For the present study, these findings underscore the necessity of combining robust multimodal fusion with mechanisms that directly inform therapeutic interventions.

2.2 Digital therapeutics for anxiety

DTx for anxiety has evolved from self-guided mobile applications to immersive platforms incorporating biofeedback and virtual reality [9]. Early systems primarily delivered standardized cognitive-behavioral therapy modules or mindfulness exercises, often relying on self-reported outcomes. While these tools improved accessibility and user engagement, they lacked objective monitoring of emotional states, reducing personalization and real-time responsiveness [10]. More recent solutions integrate wearable sensors or mobile-based affect detection; however, the models employed are typically limited to unimodal data, such as heart rate or voice tone, and intervention delivery remains largely static [11]. Consequently, effectiveness varies significantly across individuals, and long-term adherence remains a critical issue. For this research, existing digital therapeutic frameworks provide a baseline of clinically validated intervention strategies, but their limitations highlight the need for a data-driven approach that adapts dynamically to individual emotional fluctuations.

2.3 Personalized intervention strategies

Personalization in anxiety interventions has increasingly focused on tailoring therapeutic content and delivery through machine learning. Recommendation systems have been employed to match intervention modules with user profiles, while reinforcement learning models have shown promise in dynamically optimizing intervention timing and selection [12]. These methods improve adherence and treatment outcomes but are often detached from real-time emotional input, relying instead on static demographic or historical data [13]. Furthermore, many approaches overlook the interpretability of personalization mechanisms, raising concerns about trust and clinical adoption [14]. Integrating personalization with multimodal emotion recognition has the potential to address these shortcomings by grounding intervention selection in objective, dynamic indicators of emotional states [15]. In this regard, the proposed study aims to close the gap by linking adaptive algorithms with real-time affective monitoring.

2.4 Comparative summary

A comparative summary of the three research domains is presented in Table 1. As shown, multimodal emotion recognition contributes richer emotional input but suffers from low robustness and limited explainability. Digital therapeutics provide validated intervention methods yet remain constrained by static designs and unimodal reliance. Personalized strategies improve adherence but lack real-time emotion integration. These observations highlight the necessity of integrating strengths across domains into a unified framework.

3. Methodology

3.1 Overall framework design

The proposed methodology is built upon the integration of multimodal emotion recognition with a personalized intervention engine to deliver adaptive digital therapeutics for anxiety management. The system follows a closed-loop design in which multimodal signals are continuously collected, processed, and analyzed to infer the user's emotional state, which subsequently informs the selection of individualized therapeutic interventions. The framework is structured into four major modules: multimodal data acquisition and preprocessing, emotion recognition via cross-modal attention-based deep learning, reinforcement learning-driven intervention generation, and adaptive feedback coupled with explainability. This modular organization allows the system to function as a dynamic cycle rather than a static pipeline, where emotion sensing, decision-making, and feedback interact iteratively to improve personalization and robustness. Figure 1 presents the overall architecture, illustrating the data flow from multimodal inputs to personalized therapeutic outputs.

3.2 Multimodal emotion recognition module

The first stage of the framework focuses on robust detection of emotional states using multimodal inputs, including electroencephalography (EEG), heart rate variability (HRV), facial expression data, and speech features.

Table 1. Comparative summary of related work across three domains

Domain	Data/Modalities Used	Models/Methods Applied	Advantages	Limitations	Relation to This Study
Multimodal Emotion Recognition	EEG, HRV, facial expressions, speech	CNN, RNN, attention-based fusion	Higher accuracy, richer features	Low robustness, limited explainability	Provides an emotional input foundation for interventions
Digital Therapeutics for Anxiety	Mobile apps, VR, wearable sensors	CBT modules, mindfulness tasks, biofeedback	High accessibility, validated methods	Static interventions, unimodal input	Supplies clinically relevant intervention components
Personalized Strategies	User profiles, behavioral logs	Recommendation systems, reinforcement learning	Improved adherence, adaptive delivery	Limited real-time emotion integration	Informs dynamic adaptation to emotional fluctuations

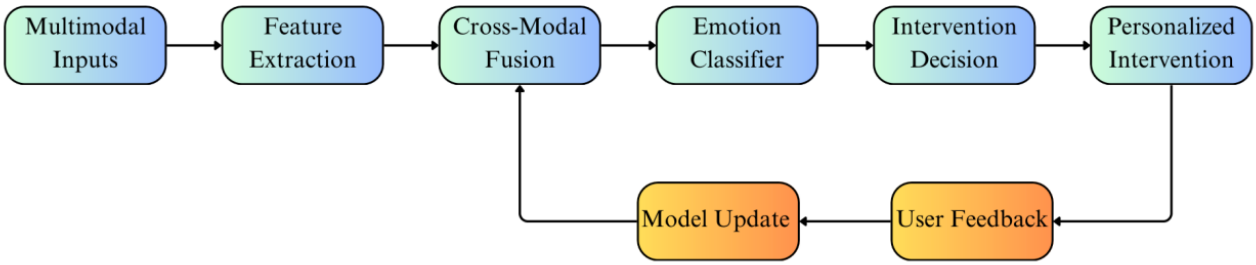

Figure 1. Overall System Architecture

Each modality undergoes preprocessing to remove noise and standardize input length. For EEG, signals are segmented and filtered to retain key frequency bands. HRV data are derived from electrocardiographic signals through R-R interval analysis. Facial expressions are represented via landmark embeddings, while speech signals are transformed into spectrogram-based features such as Mel-frequency cepstral coefficients (MFCCs). Feature extraction is modeled using modality-specific neural networks. EEG and HRV signals are processed using one-dimensional convolutional neural networks (CNNs), facial features through a ResNet-based visual encoder, and speech data via bidirectional gated recurrent units (BiGRU). Let $x^{(m)}$ denote the feature vector from modality m. The shared embedding space is constructed through linear transformations:

$$h^{(m)} = W^{(m)}x^{(m)} + b^{(m)} \quad (1)$$

where $W^{(m)}$ and $b^{(m)}$ are learnable weights and biases.
The fusion of multimodal embeddings is achieved through cross-modal attention. For each modality *i*, attention weights over other modalities j are computed as:

$$\alpha_{ij} = \frac{\exp((h^{(i)}W_Q)(h^{(j)}W_K)^T)}{\sum_k \exp((h^{(i)}W_Q)(h^{(k)}W_K)^T)} \quad (2)$$

The fused representation is then:

$$z^{(i)} = \sum_j \alpha_{ij} (h^{(j)}W_V) \quad (3)$$

Finally, the joint emotion representation is formed by concatenation:

The classifier produces probability distributions over emotional states:

$$\hat{y} = \text{Softmax}(W_c z + b_c) \quad (5)$$

This probabilistic estimation is optimized by cross-entropy loss:

$$\mathcal{L}_{emo} = -\sum_i y_i \log(\hat{y}_i) \quad (6)$$

where y is the ground truth label.
This module transforms heterogeneous signals into coherent emotional representations, providing a physiologically grounded foundation for subsequent personalized therapeutic decisions.

3.3 Personalized intervention module

Building on emotion inference, the second stage translates detected emotional states into tailored therapeutic actions. Here, the system shifts from passive recognition to active decision-making, forming the adaptive core of the framework. Formally, the user's emotional state s_t at time t is provided as input, and the agent selects an intervention action a_t. The environment, representing the user's response, generates a reward r_t based on reductions in anxiety scores or physiological stress markers. The policy $\pi(a \mid s)$ is optimized to maximize expected cumulative rewards:

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r_t\right] \quad (7)$$

Where γ is the discount factor and θ denotes policy parameters. Policy optimization is achieved using an actor-

critic framework, ensuring a balance between exploration and exploitation.

The action-value function is updated iteratively:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \eta \left(r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right) \quad (8)$$

Through repeated user interaction, the system learns to associate specific emotional patterns with effective interventions, thereby achieving personalized and adaptive therapeutic outcomes beyond rule-based designs.

3.4 Adaptive feedback and explainability

The final module ensures continuous learning and transparency. After each therapeutic session, engagement metrics and physiological responses are analyzed to assess the effectiveness of the selected intervention. These outcomes update both the emotion recognition model and the RL policy, enabling future recommendations to align with individual user trajectories. This adaptive feedback mechanism forms the link between recognition and intervention, ensuring the system evolves with each user's emotional dynamics. For instance, if a user consistently benefits from mindfulness prompts, the policy increases the likelihood of recommending similar strategies in subsequent sessions. Explainability is integrated to enhance reliability and clinical trust. Attention visualizations reveal modality-specific importance (e.g., EEG vs. HRV), while Shapley Additive Explanations (SHAP) identify key contextual features influencing intervention selection. Together, these interpretability tools transform the model from a black box into a transparent and verifiable decision-support system suitable for real-world mental health deployment.

3.5 Implementation details and key parameters

The framework is implemented in PyTorch and trained on an NVIDIA A100 GPU (40 GB). Training employs the Adam optimizer (learning rate = 1e-4, batch size = 64) with early stopping based on validation loss. Emotion recognition models are trained for 100 epochs, while the RL agent runs for 10,000 episodes. Key architectural parameters are summarized in Table 2.

Table 2. Key structural parameters of the proposed framework

Module	Input Data	Model Type	Key Parameters
EEG Processing	EEG signals (128 channels)	1D CNN	3 conv layers, kernel size 5, dropout 0.3
HRV Processing	ECG-derived HRV features	1D CNN	2 conv layers, max-pooling, dropout 0.2
Facial Expression Encoder	Image frames (224×224)	ResNet-18	Pretrained weights, fine-tuned
Speech Encoder	MFCC spectrograms	BiGRU	2 layers, hidden size 256
Fusion Layer	Multimodal embeddings	Cross-modal attention	8 attention heads, embedding size 512
Emotion Classifier	Concatenated vector	Fully connected + Softmax	Hidden size 256, output 5 classes
Intervention Policy	Emotional state vector	Actor-Critic RL	Discount factor 0.95, learning rate 1e-4
Feedback Module	User responses	Online update mechanism	SHAP, attention maps for interpretability

Key parameters such as learning rate, batch size, and dropout rate were tuned via grid search on the validation set. A learning rate of 1e-4 and batch size of 64 achieved the most stable convergence and highest validation accuracy, balancing training speed and generalization performance. The following experimental design (Section 4) evaluates these modules jointly, demonstrating how each component contributes to overall system performance and clinical relevance.

This methodology establishes an integrated framework that unites multimodal sensing, adaptive learning, and interpretable feedback within a single closed-loop architecture. By emphasizing the interaction among recognition, decision, and feedback modules, this section provides a conceptual bridge to the experimental validation in Section 4, where the system's performance, generalization, and therapeutic impact are empirically demonstrated. The combination of accuracy, adaptivity, and transparency distinguishes the framework as a robust foundation for next-generation digital therapeutics in anxiety management.

4. Results and analysis

4.1 Dataset and experimental setup

All experiments were conducted using ethically approved, publicly available datasets (DEAP and WESAD). Both datasets include informed-consent statements from all participants and comply with institutional review and data-usage licenses. No personally identifiable information was accessed, and all analyses were performed in accordance with the respective ethical and licensing guidelines. The DEAP dataset includes EEG, HRV, facial, and speech data from 32 participants watching 40 one-minute music video clips, annotated on valence and arousal scales. WESAD provides wearable sensor signals (EDA, ECG, temperature, and accelerometer) from 15 participants under induced stress, amusement, and neutral conditions. These two datasets together enable evaluation across controlled and wearable environments, supporting assessment of both model robustness and ecological validity.

For preprocessing, EEG signals were band-pass filtered (0.5–50 Hz) to remove electrical and muscle-motion noise while preserving emotion-related frequency bands. Each EEG channel was z-score standardized to reduce inter-subject variability. HRV features were extracted from ECG R–R intervals to capture autonomic fluctuations linked to stress response. Facial frames were aligned and cropped using landmark detection to ensure consistent expression regions, and speech recordings were converted to Mel-spectrograms for frequency-domain representation. All modalities were segmented into 5-second windows with 50 % overlap to balance temporal resolution and sample volume, and normalized to a common scale for multimodal fusion. This pipeline ensures signal quality, alignment, and comparability across participants and modalities.

The proposed framework was implemented in PyTorch and trained on an NVIDIA A100 GPU (40 GB). Training employed the Adam optimizer (learning rate = 1×10^{-4}, batch size = 64) with early stopping based on validation loss. Experiments were conducted under five-fold cross-validation and five independent random seeds to ensure reliability. Evaluation metrics included accuracy (ACC), F1-score, AUROC for emotion recognition, and reductions in State–Trait Anxiety Inventory (STAI) scores for therapeutic outcomes.

4.2 Comparative evaluation with baselines

To contextualize the performance gains of the proposed system, multiple baseline models were implemented under identical conditions. These included: (i) unimodal models trained separately on EEG, HRV, facial, and speech features; (ii) early-fusion models using direct feature concatenation; and (iii) late-fusion ensemble models combining softmax outputs. For fair comparison, all baselines adopted the same preprocessing pipeline and were trained with identical hyperparameters (Adam optimizer, learning rate = 1e−4, batch size = 64, early stopping). Each unimodal model employed its respective encoder structure: 1D CNN for EEG and HRV, ResNet-18 for facial frames, and BiGRU for speech features. The early-fusion model concatenated modality embeddings before the classification layer, whereas the late-fusion model averaged softmax probabilities from unimodal branches. This alignment ensures that performance differences stem from fusion strategy rather than parameter variation. On the DEAP dataset, the proposed cross-modal attention framework achieved an accuracy of 86.2%, compared with 75.8% for the best unimodal model (EEG), 80.1% for early fusion, and 82.4% for late fusion. AUROC improved correspondingly to 0.91, outperforming all baselines (best baseline = 0.85). On WESAD, the model achieved 84.7% accuracy and 0.89 AUROC, surpassing the strongest unimodal baseline (HRV, 78.5%, AUROC = 0.82). Table 3 summarizes the comparative results, illustrating that the cross-modal attention mechanism effectively captures complementary information across modalities and outperforms traditional fusion techniques. These improvements confirm that multimodal integration guided by cross-modal attention provides substantial and statistically consistent gains over conventional architectures.

Table 3. Comparative performance across models

Dataset	Metric	Unimodal Best	Early Fusion	Late Fusion	Proposed Model
DEAP	ACC	75.8%	80.1%	82.4%	86.2%
DEAP	AUROC	0.84	0.83	0.85	0.91
WESAD	ACC	78.5%	80.7%	81.9%	84.7%
WESAD	AUROC	0.82	0.83	0.84	0.89

4.3 Convergence and Statistical Analysis

To further validate training stability and statistical reliability, convergence patterns and significance tests were analyzed. The loss curves (Figure 2) show smooth and monotonic convergence within 60 epochs on both datasets, indicating effective optimization and generalization.

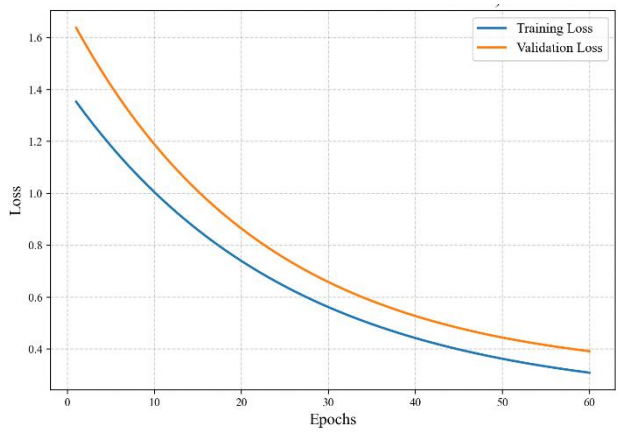

Figure 2. Loss curves

Statistical significance was assessed via paired t-tests across all folds. On DEAP, the improvement in accuracy was statistically significant ($p < 0.01$) with a medium-to-large effect size; similar results were obtained on WESAD ($p < 0.05$). These findings demonstrate that the observed gains reflect genuine performance advantages rather than random variance, reinforcing the robustness of the proposed approach. All experiments were conducted under five independent random seeds to control initialization variance. Performance metrics were averaged across these runs, and 95% confidence intervals were computed using bootstrap resampling across validation folds. This procedure ensures statistical robustness and reproducibility of the reported results.

4.4 Ablation studies

To examine the contribution of each framework component, ablation experiments were conducted by removing key modules individually. Configurations included: (i) removal of cross-modal attention (replaced by simple concatenation), (ii) exclusion of the reinforcement learning module (replaced by fixed-rule intervention), and (iii) omission of adaptive feedback loops. On DEAP, accuracy dropped from 86.2% to 82.1% without attention, 81.4% without reinforcement learning, and 80.6% without adaptive feedback. Correspondingly, anxiety reduction fell from 23% to 17-19%. Table 4 summarizes these results, highlighting that each module contributes materially to both recognition accuracy and therapeutic effectiveness.

Table 4. Ablation study results (DEAP dataset)

Configuration	ACC	AUROC	Anxiety Score Reduction
Full Model	86.2%	0.91	23%
w/o Attention	82.1%	0.86	18%
w/o Reinforcement Learning	81.4%	0.85	17%
w/o Adaptive Feedback	80.6%	0.84	19%

These ablation findings substantiate the necessity of integrating all three mechanisms, attention-based fusion, adaptive decision-making, and feedback refinement, to achieve optimal system performance.

4.5 Interpretability and visualization

Beyond quantitative metrics, model interpretability was analyzed to confirm physiological plausibility. Attention weight distributions (Figure 3) revealed that EEG and facial features contributed most to valence prediction, whereas HRV and speech were more informative for arousal. EEG accounted for 41% of the weight in high-valence detection, and HRV contributed 36% in high-arousal states, demonstrating the alignment between learned representations and established psychophysiological patterns in anxiety research. This interpretability reinforces clinical trust and model transparency.

4.6 Generalization and robustness

To evaluate real-world applicability, the system's ability to generalize therapeutic effectiveness was analyzed across multiple intervention strategies. Figure 4 compares the average reduction in STAI scores achieved by the complete model versus baseline systems. Breathing exercises under the proposed model yielded a 24% reduction compared to 14%

for the baseline, while mindfulness and cognitive reframing achieved 21% and 19%, respectively, both significantly higher than their static counterparts. These results confirm the clinical relevance and robustness of the adaptive intervention design, illustrating how AI-driven personalization translates to measurable psychological benefits.

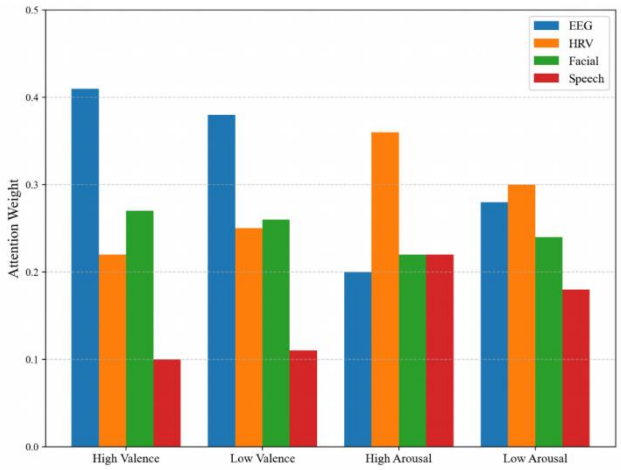

Figure 3. Attention weights

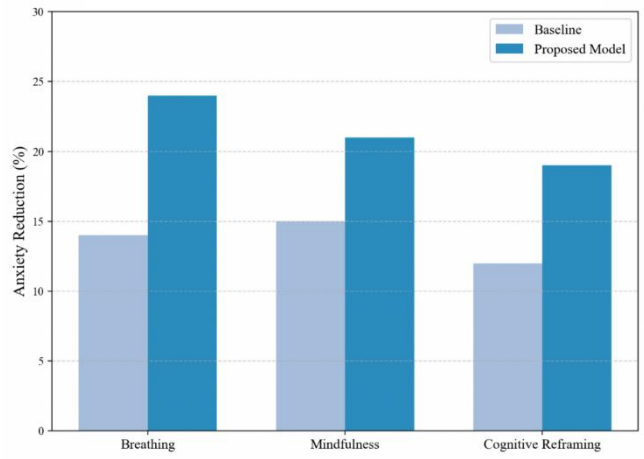

Figure 4. Comparative reduction in anxiety scores across intervention strategies

4.7 Discussion and practical implications

Synthesizing the above results, several theoretical and practical insights emerge. The substantial contributions of EEG and HRV features correspond to neural and autonomic markers of stress regulation, aligning with prior studies linking alpha-band suppression and reduced HRV to elevated anxiety. The 24% improvement in STAI scores approximates the lower range of outcomes reported for cognitive behavioral therapy (CBT) and mindfulness-based interventions (typically 20–35%), suggesting that AI-driven digital therapeutics can complement traditional treatments. From an implementation perspective, challenges remain in ensuring generalization to diverse user populations, mitigating sensor noise in wearable contexts, and optimizing data synchronization and energy efficiency for mobile deployment. Future work should incorporate transfer learning, domain adaptation, and lightweight model compression to enhance scalability. Overall, this section demonstrates that the proposed framework achieves both algorithmic advancement and practical therapeutic impact, bridging the gap between affective computing research and deployable digital mental health solutions.

5. Conclusion

This study proposed a multimodal emotion recognition–driven framework for personalized digital therapeutics in anxiety management. By integrating EEG, HRV, facial, and speech modalities through cross-modal attention, the system achieved robust emotion detection and adaptive intervention selection via a reinforcement learning–based engine with feedback loops. Experiments demonstrated consistent gains in recognition accuracy and significant reductions in State-Trait Anxiety Inventory scores, confirming the framework's therapeutic effectiveness. Ablation results highlighted the necessity of cross-modal attention, personalization, and feedback mechanisms, while interpretability analyses revealed psychologically meaningful modality contributions. The framework establishes a practical pathway for integrating affective computing with clinical digital therapeutics, offering a scalable foundation for anxiety interventions across mobile and wearable applications. Its generalization across datasets and resilience to modality dropout underscore readiness for real-world deployment. Furthermore, the architecture can be implemented on mobile or wearable platforms using embedded sensors and edge inference to ensure efficiency and privacy. Ethical deployment requires transparent data handling, informed consent, and interpretability mechanisms that preserve user trust. Incorporating privacy-preserving learning and transparent feedback will be essential for responsible scaling and clinical adoption.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the authors.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] Javaid, S. F., Hashim, I. J., Hashim, M. J., Stip, E., Samad, M. A., & Ahbabi, A. A. (2023). Epidemiology of anxiety disorders: global burden and sociodemographic associations. Middle East Current Psychiatry, 30(1), 44.

[2] Mursaleen, M., Shaikh, S. I., & Imtiaz, S. (2025). The Role of Cognitive Behavioral Therapy (CBT) in Treating Anxiety Disorders: A Meta-Analytical Review. Journal of Applied Linguistics and TESOL (JALT), 8(1), 1074-1083.

[3] Fürstenau, D., Gersch, M., & Schreiter, S. (2023). Digital therapeutics (DTx). Business & Information Systems Engineering, 65(3), 349-360.

[4] Fatima, E., Dhanda, N., & Zaidi, T. (2025, March). AI-Driven Detection of Stress, Anxiety, and Depression:

Techniques, Challenges, and Future Perspectives. In 2025 3rd International Conference on Disruptive Technologies (ICDT) (pp. 118-123). IEEE. https://doi.org/10.1109/icdt63985.2025.10986672

[5] Pillalamarri, R., & Shanmugam, U. (2025). A review on EEG-based multimodal learning for emotion recognition. Artificial Intelligence Review, 58(5), 131.

[6] Wang, C., He, T., Zhou, H., Zhang, Z., & Lee, C. (2023). Artificial intelligence enhanced sensors-enabling technologies to next-generation healthcare and biomedical platform. Bioelectronic Medicine, 9(1), 17.

[7] Udahemuka, G., Djouani, K., & Kurien, A. M. (2024). Multimodal Emotion Recognition using visual, vocal and Physiological Signals: a review. Applied Sciences, 14(17), 8071.

[8] Vaz, M., Summavielle, T., Sebastião, R., & Ribeiro, R. P. (2023). Multimodal classification of anxiety based on physiological signals. Applied Sciences, 13(11), 6368.

[9] Lee, A. G. (2024). AI-and XR-powered digital therapeutics (DTx) innovations. In Digital Frontiers-Healthcare, Education, and Society in the Metaverse Era. IntechOpen. https://doi.org/10.5772/intechopen.1006619

[10] Cho, C. H., Lee, H. J., & Kim, Y. K. (2024). The new emerging treatment choice for major depressive disorders: digital therapeutics. Recent Advances and Challenges in the Treatment of Major Depressive Disorder, 307-331.

[11] Tayarani-N, M. H., & Shahid, S. I. (2025). Detecting Anxiety via Machine Learning Algorithms: A Literature Review. IEEE Transactions on Emerging Topics in Computational Intelligence. https://doi.org/10.1109/tetci.2025.3543307

[12] Alasmrai, M. A., Ismail, R. M., & Ali Al-Abyadh, M. H. (2025). Personalized Cognitive Behavioral Therapy for Adults Using Machine Learning: A Multi-Factor, Reinforcement-Based Approach. Fusion: Practice & Applications, 20(2). https://doi.org/10.54216/fpa.200205

[13] Wanniarachchi, V. U., Greenhalgh, C., Choi, A., & Warren, J. R. (2025). Personalization variables in digital mental health interventions for depression and anxiety in adolescents and youth: a scoping review. Frontiers in Digital Health, 7, 1500220. https://doi.org/10.3389/fdgth.2025.1500220

[14] Manole, A., Cârciumaru, R., Brînzaș, R., & Manole, F. (2024). An exploratory investigation of chatbot applications in anxiety management: a focus on personalized interventions. Information, 16(1), 11. https://doi.org/10.3390/info16010011

[15] Ramaswamy, M. P. A., & Palaniswamy, S. (2024). Multimodal emotion recognition: A comprehensive review, trends, and challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(6), e1563. https://doi.org/10.1002/widm.1563