Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 3 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i03.60249

PAPER

Mobile Music Therapy Integrating AI-Driven Emotion Prediction and a Human-Computer Interaction Experience Model

Ruqi Bai (✉)

XinZhou Normal University, Xinzhou, China

18603502125@163.com

ABSTRACT

The integration of affective computing with mobile health technologies has created transformative opportunities for scalable and personalized music therapy. However, traditional music therapy remains limited by insufficient personalization, subjective emotion assessment, and low adaptability in human–computer interaction (HCI). Existing artificial intelligence (AI) emotion prediction methods also lack robust multimodal data fusion in mobile environments and alignment with clinical workflows. A closed-loop mobile music therapy system incorporating multimodal AI emotion prediction was developed in this study to address these limitations. The system integrates multimodal emotion sensing, dynamic therapy matching, empathic interaction, and closed-loop optimization. A convolutional neural network-long short-term memory (CNN-LSTM) hybrid model was employed to predict emotional states using physiological signals, behavioral features, and subjective feedback, while a prescriptive dynamic music therapy model and a multidimensional HCI evaluation framework were constructed. Experimental results showed that the multimodal CNN–LSTM model outperformed unimodal models and traditional algorithms, and the closed-loop system achieved significantly greater improvements in emotional regulation and stress reduction than non-adaptive interventions (traditional therapy approaches and static playlists). Dynamic enhancements in heart rate variability and reductions in cortisol levels provided objective physiological evidence of therapeutic efficacy. The empathic interaction framework increased usability and treatment adherence. Mediation analysis confirmed the pathway linking emotion prediction, music therapy matching, and outcome enhancement. Superior efficacy in individuals with anxiety or depressive tendencies further highlighted scenario-specific therapeutic responsiveness compared with postoperative rehabilitation groups. This study advances the integration of AI and music therapy, and the proposed real-time mapping model linking emotional states, music parameters, and interaction feedback provides an actionable framework for the development of empathic AI systems in healthcare. The closed-loop design establishes a new paradigm for personalized mobile medical interventions with significant clinical and interdisciplinary relevance.

Bai, R. (2026). Mobile Music Therapy Integrating AI-Driven Emotion Prediction and a Human-Computer Interaction Experience Model. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(3), pp. 55–70. https://doi.org/10.3991/ijim.v20i03.60249

Article submitted 2025-10-13. Revision uploaded 2025-11-22. Final acceptance 2025-11-25.

KEYWORDS
multimodal AI emotion prediction, closed-loop mobile music therapy, personalized clinical intervention, empathic HCI, therapeutic mechanism

1 INTRODUCTION

Music therapy, an interdisciplinary field integrating musicology, psychology, and clinical medicine [1, 2], has been widely validated for its clinical efficacy in emotional regulation, chronic pain management, and neurological rehabilitation. However, the global scarcity of registered music therapists [3, 4]—currently estimated at only 1.2 therapists per 100,000 people—has made it difficult for traditional treatment models to meet large-scale service demands. Consequently, digital transformation has emerged as an essential pathway for overcoming structural limitations within the field. The widespread adoption of wearable devices and breakthroughs in affective computing have enabled music therapy practices to expand beyond professional clinical settings into everyday environments [5, 6]. These technological advancements provide robust foundations for real-time emotion sensing and personalized therapeutic intervention, while also creating new opportunities for interdisciplinary research. Nevertheless, traditional music therapy continues to be constrained by inherent limitations, including heavy reliance on practitioner expertise, insufficient personalization, and a restricted service radius [7, 8]. Existing AI-driven health applications, in turn, lack closed-loop alignment with clinical treatment workflows [9], preventing them from simultaneously achieving three essential requirements: clinical efficacy, real-time personalization, and high treatment adherence. This "triple-gap paradox" has become a central barrier to the clinical implementation of AI-assisted therapeutic systems.

A review of international research trends reveals several persistent shortcomings. AI-based emotion prediction studies have predominantly focused on unimodal data sources, resulting in inadequate robustness under mobile conditions [10]. Research on digital music therapy has largely relied on static playlists or rule-based recommendation mechanisms, neither of which accommodates real-time emotional fluctuations [11, 12]. Mobile health HCI research has emphasized usability optimization while overlooking the fundamental relationship between empathic experiences and treatment adherence [13, 14]. A comprehensive gap persists in existing research, as no integrated framework has yet been established to concurrently address three core requirements: consistency between clinical efficacy and therapeutic logic, real-time personalization with adaptivity to dynamic emotional states, and high user adherence supported by empathic, low–cognitive load interaction [15, 16]. The present study aims to fill this interdisciplinary gap through the development of a closed-loop system.

The core objective of this study is to design a closed-loop mobile music therapy system, which integrates multimodal AI emotion prediction, to construct a scientifically grounded HCI experience model, and to systematically evaluate its clinical efficacy, personalized adaptability, and user adherence. To achieve this objective, three fundamental research questions were addressed. First, how multimodal data fusion can be employed to enhance the accuracy and robustness of emotion prediction under mobile conditions was examined. Second, how a prescriptive,

dynamic music therapy model driven by real-time emotional feedback can be constructed to enable genuine personalized intervention was explored. Third, how an empathic HCI framework can be designed to establish a dynamic balance among usability, affective resonance, and treatment adherence was investigated. The structure of this study is arranged below. Section 2 presents the system architecture in detail, including the implementation of the emotion prediction algorithm, the therapeutic model, and the HCI interface. Section 3 introduces the multidimensional HCI evaluation system. Section 4 reports the experimental validation procedures and findings. Section 5 summarizes the significance of the study and outlines future research directions.

2 SYSTEM DESIGN

2.1 Overall architecture

The proposed system was designed using a three-layer cross-hierarchical collaborative architecture, in which data-driven processing and closed-loop iteration were adopted as core principles. This architecture enables the deep integration of emotion sensing, therapeutic intervention, and interaction feedback, while the modular design ensures component independence and compatibility, supporting the fully automated implementation of personalized music therapy workflows. The data layer is responsible for the acquisition and preprocessing of multimodal emotional data. Physiological signals, behavioral features, and subjective feedback are synchronously collected through the integration of wearable devices and mobile terminals. Environmental noise is removed using denoising algorithms, and data scales are unified using Z-score normalization, expressed as $x' = (x-\mu)/\sigma$, where μ denotes the mean and σ represents the standard deviation. The processed, high-quality data are used for subsequent processes. The algorithm layer functions as the system's central decision unit, establishing the closed-loop coupling between the emotion prediction model and the dynamic therapeutic model. Emotion prediction is performed using multimodal data to generate real-time emotion assessments, while the dynamic therapeutic model produces personalized intervention prescriptions and is continuously refined through the return of feedback data. The application layer adopts an empathic design paradigm to construct the HCI interface. This layer integrates emotion assessment, therapeutic session delivery, and progress tracking, enabling treatment management and feedback collection while minimizing cognitive load.

The core operational workflow follows a closed-loop iterative logic. After acquisition and preprocessing, multimodal data are transferred to the algorithm layer for emotional state inference. The inferred emotional state then drives the dynamic therapeutic model, which generates a music therapy prescription that is presented through the application layer. User interaction behaviors and subjective feedback are transmitted back in real time to refine the parameters of the emotion prediction model and guide dynamic adjustments to the therapeutic scheme. Through continuous iteration across the steps of acquisition, prediction, matching, feedback, and optimization, the precision and adaptive capacity of the intervention are progressively enhanced. Figure 1 presents the logical architecture of the mobile music therapy system integrating multimodal emotion prediction.

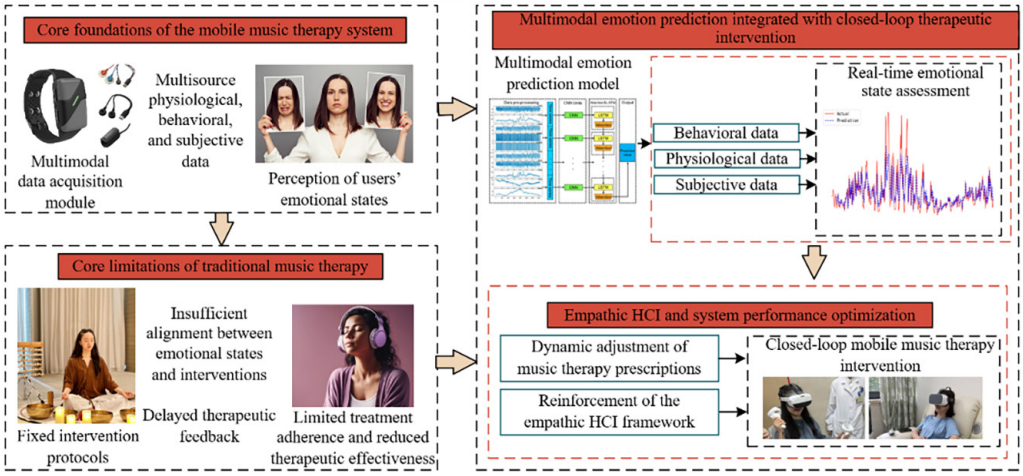

Fig. 1. Logical architecture of the mobile music therapy system integrating multimodal emotion prediction

2.2 Design of the multimodal emotion prediction algorithm

The multimodal emotion prediction algorithm was designed to enhance the accuracy and robustness of emotional state assessment in mobile environments. By integrating complementary information from physiological, behavioral, and subjective data, a technical pipeline of data diversity, feature fusion, and model optimization was established to accommodate the resource constraints characteristic of mobile environments.

A multi-source complementary strategy was adopted for data acquisition. Wearable devices were used to collect heart rate variability and electrodermal activity, with signal stability improved through optimized denoising algorithms. Mobile terminals extracted facial micro-expression features using OpenCV, and voice features were obtained using Mel-Frequency Cepstral Coefficients (MFCCs). The MFCCs are computed as:

$$MFCC_n = \sum_{k=1}^{K-1} |X(k)|^2 H_n(k) \tag{1}$$

where, $X(k)$ denotes the spectral coefficient, and $H_n(k)$ represents the Mel filterbank. A lightweight emotion rating scale was used to rapidly collect subjective feedback, forming a multidimensional emotional dataset.

Feature engineering incorporates an attention mechanism to achieve multimodal fusion by dynamically amplifying the contribution of highly relevant emotional cues. The weights are calculated as:

$$a_i = \frac{\exp(s_i)}{\sum_{j=1}^{M} \exp(s_j)} \tag{2}$$

where, a_i denotes the attention weight for the i-th modality, s_i represents the modality importance score, and M is the total number of modalities. A CNN–LSTM hybrid architecture was employed as the predictive model. The CNN component extracts local salient patterns from the fused feature space, while the LSTM

component captures temporal emotional dynamics. Together, the two components support the simultaneous modeling of static features and temporal trends. Experimental validation indicated that an accuracy of 89.7% was achieved, significantly surpassing traditional algorithms such as support vector machines (SVMs) and Random Forests (RFs). The lightweight design ensures real-time operational performance on mobile devices, providing reliable decision support for closed-loop therapy.

2.3 Design of the dynamic music therapy model

The dynamic music therapy model was designed with clinical efficacy and personalized adaptability as its core objectives. By defining therapeutic goals, constructing a standardized music resource library, and designing a closed-loop matching mechanism, precise alignment was achieved between emotion assessment and intervention implementation. This ensures continuous adaptation of the intervention scheme to users' emotional states, clinical requirements, and individual characteristics. The model design follows a principle of algorithmic translation of clinical logic, in which professional music therapy expertise is transformed into quantifiable, executable rules.

Therapeutic goals were categorized into three primary scenarios: emotional relief, pain management, and sleep improvement. Corresponding outcome indicators were clearly defined for each scenario. Emotional relief is evaluated by reductions in anxiety and depression scale scores; pain management effectiveness is quantified using the visual analog scale (VAS); and sleep improvement is assessed using the sleep quality index. The music resource library was constructed based on a two-dimensional valence–arousal model, in which each music item was assigned a quantified coordinate (V,A). Valence (V) ranges from –1 to 1, and arousal (A) also ranges from –1 to 1. Based on their distribution in the valence–arousal space, music items were further annotated with therapeutic attributes: relaxation (low A, medium-to-high V), activation (high A, medium-to-high V), and soothing (low A, medium V), yielding a structured and searchable music resource system.

A prescriptive dynamic matching mechanism was implemented through a three-stage strategy comprising initial matching, real-time adjustment, and generative supplementation. Initial matching was performed using baseline emotional assessments, individual characteristics, and clinical requirements. A weighted scoring formula was applied to compute the suitability of each music item, producing the initial intervention scheme:

$$S = w_E E + w_F F + w_C C \tag{3}$$

where, E denotes the baseline emotional value, F represents the individual characteristic vector, C indicates the clinical requirement weight, and w_E, w_F, and w_C are the normalized weighting coefficients.

During the real-time adjustment stage, dynamic modifications to musical tempo, volume, and instrumental composition were performed through a parameter optimization model according to fluctuations in the predicted emotional state. For example, in anxiety-related conditions, synchronizing music was first applied to stabilize

respiratory rhythms, followed by a gradual transition toward low-arousal music using an arousal adjustment formula:

$$A_{new} = A_{current} - k\Delta E \tag{4}$$

where, k denotes the adjustment coefficient, and ΔE represents the degree of emotional improvement.

A generative supplementation module was incorporated to address emotional states that are insufficiently covered by the traditional music resource library. This module employs generative adversarial networks (GANs) and variational auto encoders (VAEs) to construct a music generation model capable of producing custom therapeutic music based on the two-dimensional coordinates of the target emotional state, thereby filling gaps in resource coverage. An explainability module was integrated to translate the recommendation logic into interpretable natural language descriptions based on the model's decision pathways. By linking emotional features with the physiological and psychological mechanisms underlying music-based intervention, the explainability module can enhance user trust and acceptance of the intervention scheme.

2.4 Design of empathic HCI framework

The empathic HCI framework was designed with the core principles of reducing cognitive load, enhancing emotional resonance, and ensuring data security. Through multidimensional design encompassing interface optimization, process simplification, feedback enhancement, and privacy protection, a human-centered interaction experience suitable for medical environments was constructed, thereby improving treatment adherence and emotional acceptance. Differences in user abilities across populations were fully considered to achieve a balance between usability and professionalism.

Interface design follows a minimalist principle, adopting simplified layouts and high-contrast visual elements to reduce informational distractions. Responsive design was implemented to accommodate variations in screen size, meeting the needs of elderly users and individuals with limited mobility. The interaction process was structured around a "one-touch initiation–automated intervention–intelligent feedback" logic, minimizing operational steps and reducing therapeutic initiation to a single interaction action. Voice control functionality was integrated to support natural language command recognition and execution, providing an accessible entry point for users with limited operational capability. By anticipating user needs, the process design enables automated linkage among data acquisition, emotion assessment, and intervention initiation, shortening the time from need recognition to therapeutic execution and improving overall interaction efficiency.

A multimodal coordination strategy was adopted for the feedback mechanism, integrating visual cues, voice guidance, and haptic responses to provide continuous feedback throughout the entire intervention cycle. Visual cues present treatment progress through progress bars and status icons; voice guidance delivers gentle instructions for breathing synchronization and emotional regulation; and haptic feedback provides tactile prompts at key nodes, strengthening the emotional connection between users and the system. Elements of positive psychology were

incorporated through three functional modules: growth recording, progress visualization, and achievement reinforcement. Progress visualization uses line charts to dynamically display emotional improvement trends and treatment adherence, while the achievement reinforcement module assigns tiered badges based on metrics such as total treatment duration and days of continuous engagement. User intrinsic motivation is enhanced through a self-efficacy reinforcement function expressed as:

$$SE = \alpha SE_{prior} + \beta P \tag{5}$$

where, α and β denote weighting coefficients, SE_{prior} represents prior self-efficacy, and P indicates progress completion rate.

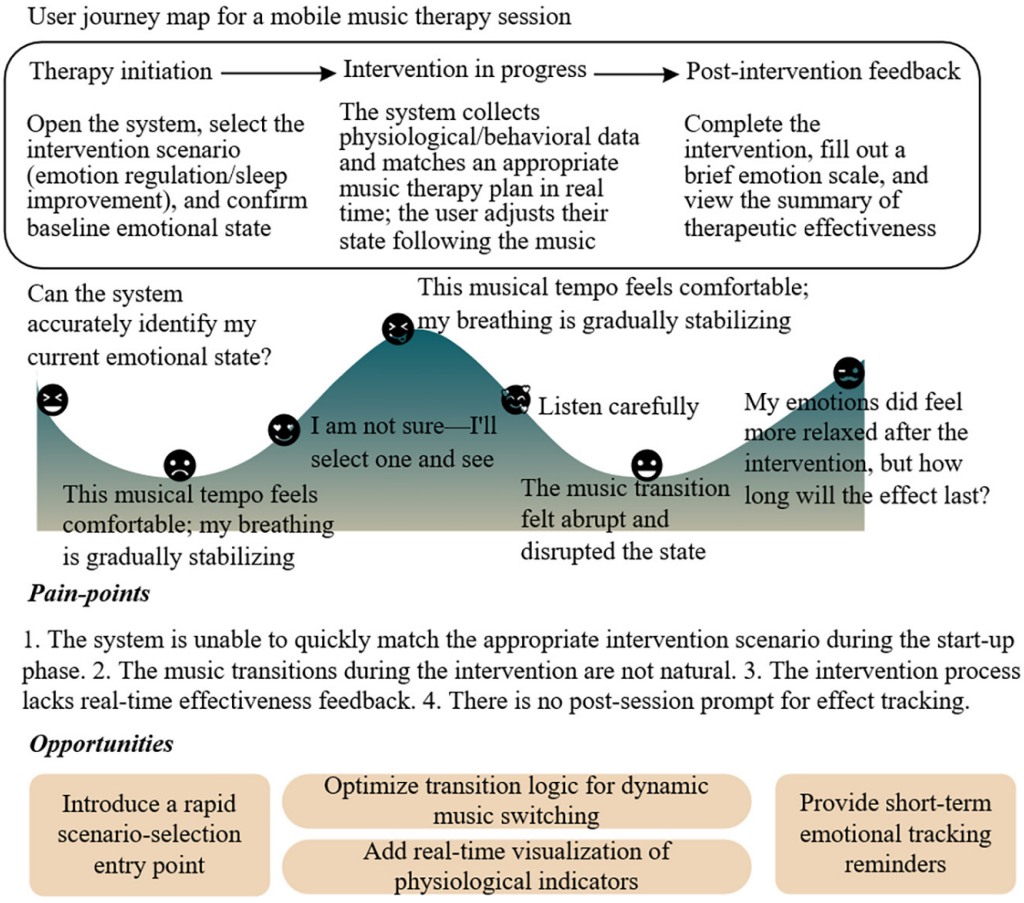

Fig. 2. User journey map for a mobile music therapy session

Privacy protection design strictly adheres to general data protection regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) standards. End-to-end encryption of sensitive health data was implemented using the Advanced Encryption Standard (256-bit key length), i.e., AES-256, and data anonymization was applied to remove personally identifiable information. A hierarchical access control system was established to ensure compliance in data usage and access, thereby reducing privacy concerns among users.

Figure 2 presents the user journey map for a single mobile music therapy intervention. It illustrates the core process of therapy initiation, intervention in progress,

and post-intervention feedback, along with users' key behavioral patterns and real-time emotional experiences. The corresponding pain points and opportunities for system optimization across each stage are also summarized.

3 INTERACTION EXPERIENCE EVALUATION SYSTEM

3.1 Evaluation index system

The interaction experience evaluation system was constructed based on the principles of multidimensionality and quantifiability. It encompasses four major dimensions—usability, emotional adaptability, clinical adaptability, and technical performance—forming a comprehensive evaluation framework that integrates system functionality, user experience, and clinical value, thereby ensuring scientific validity and practical applicability of the assessment results. Standardized quantitative methods were adopted for all indicators to facilitate data collection and cross-sectional comparison.

Usability indicators focus on system ease of use and operational efficiency, including task completion time, error rate, and System Usability Scale (SUS) scores. Task completion time is defined as the total duration from system initiation to completion of the assigned therapeutic task. Error rate is calculated as the ratio of incorrect operations to total operations. The SUS adopts a 10-item Likert five-point scale. The final score is calculated using the scoring formula shown below, yielding a total score ranging from 0 to 100, with higher scores indicating better usability.

$$SUS = 2.5 \times \sum_{i=1}^{10} x_i - 10 \tag{6}$$

The emotional adaptability indicators include emotion–music matching accuracy and emotional resonance score. Matching accuracy is calculated using the formula shown below. The emotional resonance score is obtained using a Self-Assessment Manikin (SAM) scale, which quantifies valence, arousal, and dominance across seven levels.

$$Acc = \frac{N_{match}}{N_{total}} \times 100\% \tag{7}$$

where, N_{match} denotes the number of successful matches, and N_{total} represents the total number of recommendations.

Clinical adaptability indicators evaluate the practical application value of the system, comprising treatment adherence and context adaptability. Treatment adherence is defined as the ratio of actual completed treatments to planned treatments. Context adaptability is quantified by the success rate of system use across representative environments such as the home and community settings. Technical performance indicators address the operational requirements of mobile environments and include three core indicators of emotion prediction accuracy, system response latency, and resource utilization. Emotion prediction accuracy is computed as the proportion of correctly predicted emotional samples relative to the total number of samples. System response latency is required to remain below 500 ms, and CPU and memory utilization are maintained under 20%, ensuring smooth operation on mobile devices.

3.2 Evaluation method design

A mixed-methods evaluation design combining quantitative and qualitative approaches was adopted. Objective data were obtained through controlled experiments, while subjective feedback was collected through semi-structured interviews and behavioral observation. The integration of quantitative results with qualitative insights enables mutual validation and enhances the credibility and depth of the evaluation findings.

Quantitative evaluation was conducted using a controlled experimental design comprising four parallel groups: participants in the experimental group use the closed-loop mobile music therapy system developed in this study; participants in the first control group receive traditional music therapy delivered through a fixed therapeutic protocol designed by a professional therapist; participants in the second control group use a static playlist–based mobile application without dynamic adjustment functions; and participants in the placebo group listen to random music without therapeutic properties. Data collection encompasses physiological indicators, psychological scales, and behavioral data. Physiological indicators include heart rate variability and cortisol levels, acquired through wearable devices to reflect autonomic nervous function and stress response states. Psychological assessment includes anxiety and depression scales and a music therapy outcome scale to quantify emotional improvement and therapeutic effectiveness. Behavioral data are recorded through system logs, documenting usage frequency, operation pathways, and task completion outcomes for analysis of user behavioral characteristics. All quantitative data were analyzed using repeated-measures Analysis of Variance (ANOVA) and independent-sample t-tests, with a significance threshold of $\alpha = 0.05$.

Qualitative evaluation was conducted through semi-structured interviews and behavioral observation. Semi-structured interviews were designed around core questions addressing system effectiveness, usability, and emotional experience. One-on-one in-depth interviews were carried out with 20 participants in each group, and the entire interview process was audio-recorded and transcribed. A thematic analysis approach was used to extract key perspectives. Behavioral observation was performed in naturalistic usage contexts, documenting users' operational habits, emotional reactions, and encountered difficulties. A behavioral coding framework was applied to convert observational findings into structured data. Combined with interview insights, this approach enables an in-depth analysis of the critical pain points and advantages of the overall user experience.

4 EXPERIMENTAL VALIDATION AND RESULTS ANALYSIS

4.1 Experimental design and data processing

The experiment was conducted to systematically validate the clinical efficacy, user adherence, and HCI experience of the closed-loop mobile music therapy system. A randomized controlled design was adopted to ensure the scientific rigor and reliability of the findings. A total of 120 participants were recruited, comprising 40 individuals with anxiety or depressive tendencies, 40 postoperative rehabilitation patients, and 40 healthy controls. Participants were allocated to the closed-loop system group, traditional therapy group, static-playlist

group, and random-music group using stratified random sampling, resulting in 30 participants per group. Baseline comparability across gender, age, and severity of condition was ensured. Sample size was determined using statistical power analysis with $\alpha = 0.05$ and a power of 0.8. Pre-experimental estimates indicated a minimum requirement of 26 participants per group; this number was increased to 30 to enhance robustness. Experimental equipment included iOS and Android mobile terminals, wearable devices capable of acquiring heart rate variability and electrodermal activity, and a data collection server with encrypted-storage functionality. Independent variables were defined as the four intervention modalities. Dependent variables included emotional improvement (assessed primarily via anxiety and depression scale scores), user adherence (measured through continuous usage rates), and HCI experience (measured using the SUS). Controlled variables included a uniform daily intervention duration of 30 minutes, a fixed four-week experimental period, and standardized evaluation tools and data-collection protocols. Interference factors were strictly controlled to ensure comparability across groups.

Data acquisition followed a multidimensional and real-time design, encompassing four categories: physiological data, psychological scale data, behavioral data, and system performance data. Physiological data—including heart rate variability, electroencephalographic α/β power, and cortisol levels—were continuously collected via wearable devices to reflect physiological emotional states and stress responses. Psychological scale data were obtained periodically through standardized instruments to quantify emotional status and therapeutic outcomes. Behavioral data—usage frequency, operation pathways, and task completion—were automatically logged by the system. System performance data focused on key technical indicators such as emotion prediction accuracy and response latency to evaluate operational stability. Data preprocessing followed established academic standards. Outliers were removed using the interquartile range method. Z-score normalization was applied to unify data scales, and moving-average smoothing was used for temporal data to eliminate random disturbances and noise arising from mobile environments.

4.2 Results analysis

Figure 3 presents the comparative performance of different models on the emotion prediction task. The CNN–LSTM multimodal model developed in this study demonstrates superior performance across all metrics, achieving an accuracy of 89.7%, a recall rate of 87.6%, and an F1-score of 0.886. These values are markedly higher than those obtained by the three unimodal models—physiological signal–based, behavioral feature–based, and subjective feedback–based—as well as by the traditional machine learning models, RF and SVM. Unimodal models exhibit limited performance due to reliance on single-dimensional data, with the subjective feedback modality performing the worst, attaining an accuracy of only 69.3%. Although traditional algorithms outperform unimodal models, they do not achieve the deeper level of multimodal feature integration. The CNN-LSTM multimodal model achieves performance improvements of approximately eight percentage points over the RF model, validating the technical advantages of multimodal feature fusion and the hybrid CNN-LSTM architecture in mobile emotion prediction scenarios.

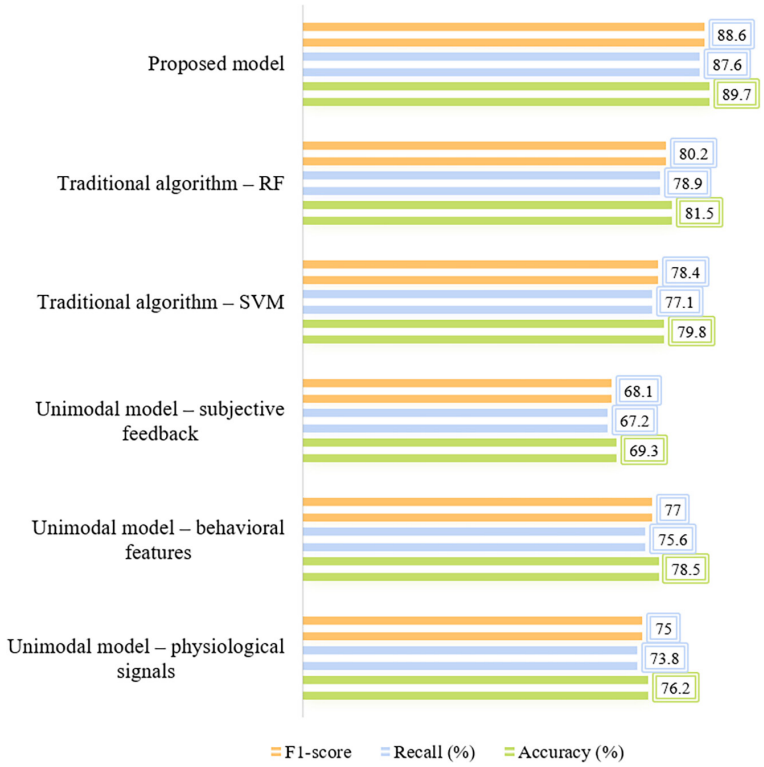

Fig. 3. Performance comparison of various emotion prediction models

Figure 4 presents the changes in core clinical indicators following intervention in each group. In the experimental group, the reductions in clinical scores were the most pronounced: the change in the Self-Rating Anxiety Scale (SAS) score was –12.6, the change in the Self-Rating Depression Scale (SDS) score was –11.8, and the reduction in cortisol level reached –1.8 μg/dL. These improvements were significantly greater than those observed in the traditional therapy group, the static-playlist group, and the placebo group, with all between-group comparisons yielding $P < 0.05$. These results indicate that the closed-loop dynamic music therapy system produces substantially stronger effects in emotional regulation and stress alleviation compared with non-adaptive intervention methods, thereby validating the clinical benefits of the dynamic matching mechanism.

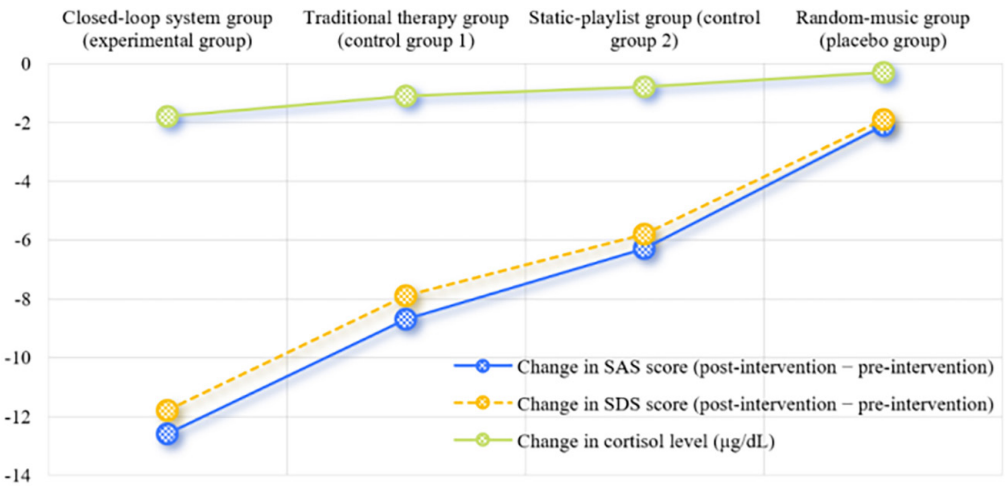

Fig. 4. Comparison of clinical indicator changes across intervention groups

Table 1. Comparison of therapeutic outcomes across subgroups

Subgroup Type	SAS Reduction Rate (%)	SDS Reduction Rate (%)	Adherence Rate (%)	P-Value (vs. Postoperative Rehabilitation Group)
Anxiety/depression-prone group	32.6	30.8	85.3	0.031
Postoperative rehabilitation group	24.1	22.5	78.6	–
Healthy control group	8.3	7.5	72.4	0.006

Table 1 shows that the anxiety/depression-prone subgroup exhibits the largest therapeutic gains, with a 32.6% reduction in SAS scores, a 30.8% reduction in SDS scores, and an adherence rate of 85.3%, all significantly higher than those observed in the postoperative rehabilitation subgroup (P = 0.031). The healthy control group demonstrates the smallest degree of improvement across all indicators. The observed differences are attributed to the distinct mechanisms underlying emotional disorders and pain-related symptoms. Music therapy interventions exert more direct regulatory effects on emotional dysregulation, whereas postoperative rehabilitation requires concurrent management of physiological pain, limiting the effectiveness of music therapy as a standalone intervention.

Figure 5 illustrates the physiological adjustments in the experimental group observed over the four-week intervention period. Heart rate variability, measured using the Root Mean Square of Successive Differences (RMSSD) index, increased steadily from 42.3 ms to 63.7 ms, indicating improved autonomic nervous system regulation. Cortisol concentration decreased from 8.7 μg/dL to 6.9 μg/dL, reflecting a reduction in physiological stress responses. In addition, the electroencephalographic α/β power ratio increased from 1.1 to 1.8, signifying reduced emotional arousal and an enhanced state of relaxation. These dynamic physiological changes provide objective biological evidence supporting the clinical effectiveness of the closed-loop music therapy system.

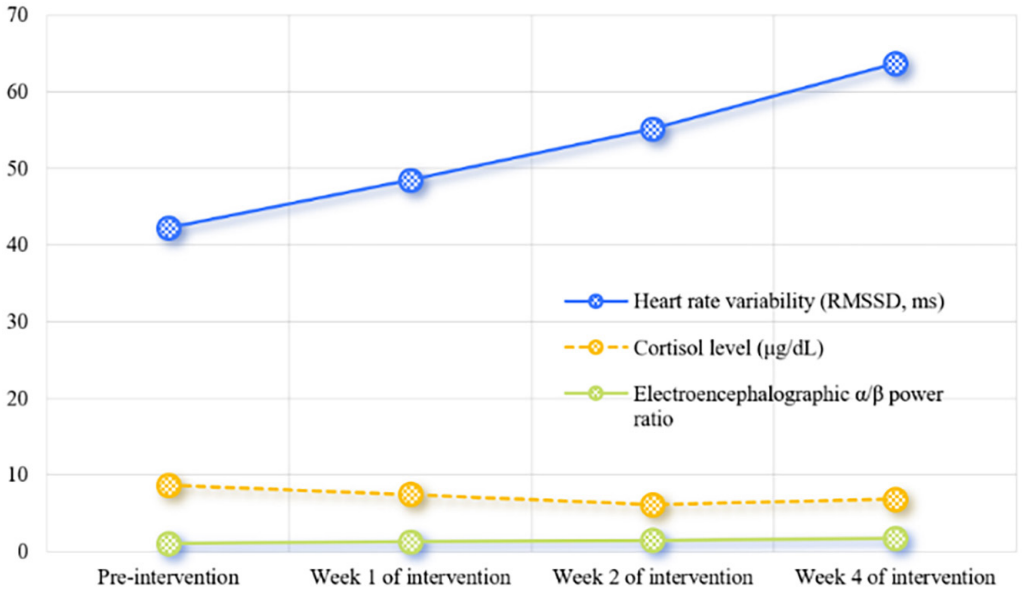

Fig. 5. Dynamic changes in physiological indicators in the experimental group during the intervention

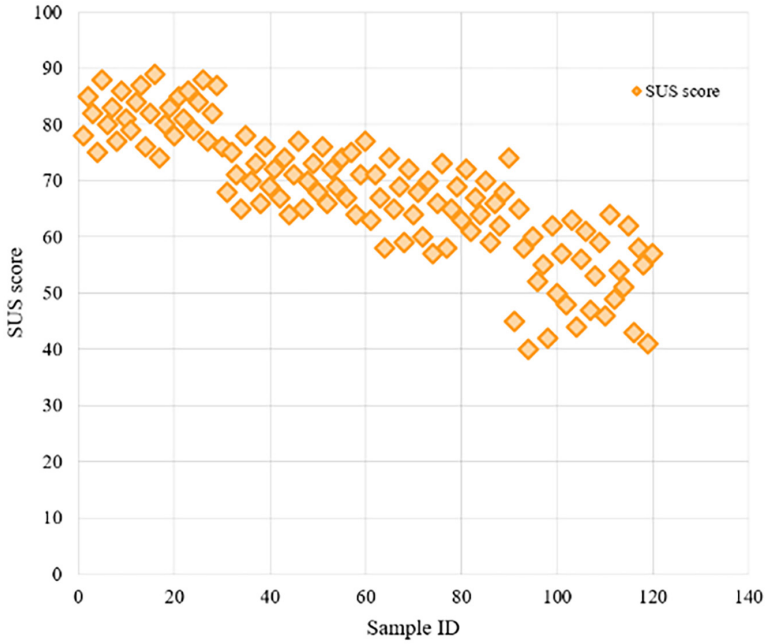

Fig. 6. Distribution of SUS scores across intervention groups

The distributions presented in Figures 6 and 7 indicate that SUS scores in the experimental group are predominantly concentrated within the 70–90 range, substantially higher than the distributions observed in the traditional therapy group (60–80), the static-playlist group (50–70), and the placebo group (40–60). Correspondingly, adherence rate samples in the experimental group are concentrated within the 70%–90% interval, whereas participants in the other three groups largely exhibit adherence rates within the 30%–70% range. These results demonstrated that the empathic HCI framework markedly enhanced system usability. Furthermore, the advantages in interaction experience were directly translated into higher treatment adherence.

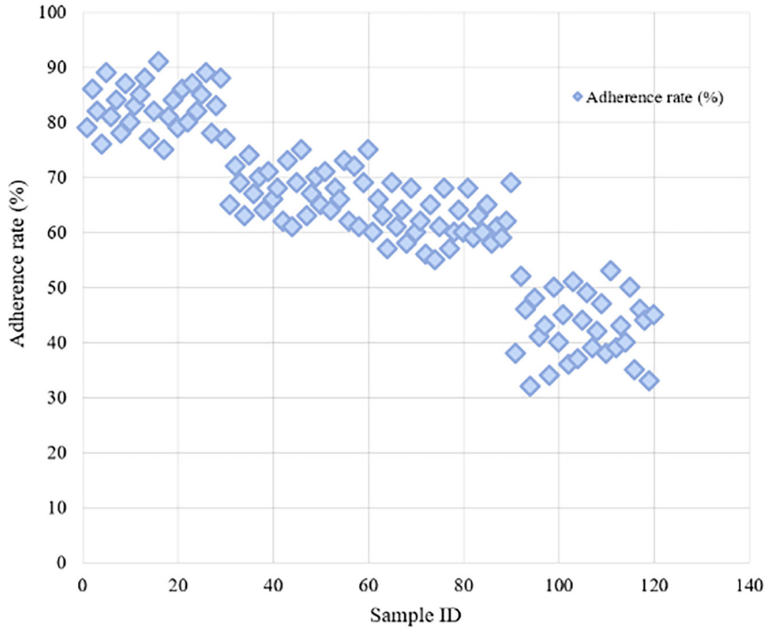

Fig. 7. Distribution of adherence rates across intervention groups

Table 2. Mediation analysis results

Pathway Relationship	Regression Coefficient (β)	Standard Error (SE)	95% Confidence Interval	P-Value
Emotion prediction accuracy → Music matching precision	0.682	0.073	[0.538, 0.826]	<0.001
Music matching precision → Emotional improvement	0.325	0.065	[0.200, 0.450]	<0.001
Emotion prediction accuracy → Emotional improvement (direct effect)	0.218	0.071	[0.080, 0.356]	0.002
Emotion prediction accuracy → Emotional improvement (mediated effect)	0.222	0.048	[0.120, 0.340]	<0.001

Table 2 shows that the direct effect of emotion prediction accuracy on emotional improvement is 0.218 (P = 0.002), whereas the mediated effect through music matching precision is 0.222, with a 95% confidence interval of [0.120, 0.340] (P < 0.001). These findings indicate that the accuracy of emotion prediction must be translated through enhanced matching precision in the music intervention scheme to effectively yield improvements in emotional outcomes. This clarifies the functional pathway of emotion prediction → music matching → therapeutic enhancement and provides theoretical support for the efficacy mechanism of the system.

5 CONCLUSION

This study addressed the central limitations of traditional music therapy—namely, the inability to scale services effectively and the lack of clinical adaptability in existing AI health applications—by designing and validating a closed-loop mobile music therapy system integrating a multimodal CNN-LSTM emotion prediction model and an empathic HCI framework tailored for medical contexts. The findings demonstrated that the proposed multimodal emotion prediction model achieved substantially higher accuracy, recall, and overall performance in mobile environments compared with unimodal models and conventional machine learning algorithms. The closed-loop therapeutic system also produced markedly superior outcomes in emotional regulation and stress reduction relative to non-adaptive intervention methods, with the most pronounced benefits observed among individuals with anxiety or depressive tendencies. Dynamic improvements in physiological indicators such as heart rate variability and cortisol levels provide objective biological evidence of therapeutic effectiveness, while the empathic interaction framework significantly enhanced system usability and treatment adherence. Mediation analysis further elucidates the functional pathway—emotion prediction → music matching → outcome enhancement. This study fills a critical theoretical and technical gap in the interdisciplinary integration of AI, music therapy, and HCI. Through a standardized system architecture and comprehensive evaluation framework, a scalable paradigm for the digital clinical translation of music therapy is established, offering practical value for alleviating the global shortage of certified music therapists.

6 REFERENCES

[1] R. T. Sampaio, "The semiotic music therapy song analysis protocol and its use as music therapy assessment," *Musica Hodie*, vol. 18, no. 2, pp. 307–326, 2018. https://doi.org/10.5216/mh.v18i2.51763

[2] S. L. Curtis, "Music therapy and social justice: A personal journey," *The Arts in Psychotherapy*, vol. 39, no. 3, pp. 209–213, 2012. https://doi.org/10.1016/j.aip.2011.12.004

[3] S. Crabtree, M. H. Hsu, J. Pool, and H. Odell-Miller, "Music therapist's use of singing, listening, playing instruments and movement with music to improve cognition and reduce neuropsychiatric symptoms in music therapy for people living with dementia," *British Journal of Music Therapy*, vol. 39, no. 1, pp. 5–19, 2025. https://doi.org/10.1177/13594575251327948

[4] A. Meadows, A. Turry, E. Schwartz, C. Fisher, and B. Matney, "Defining music therapy musicianship: An analysis of music therapists' clinical work," *Journal of Music Therapy*, vol. 62, no. 1, 2025. https://doi.org/10.1093/jmt/thaf002

[5] A. H. D. Crooke and K. S. Mcferran, "Improvising using beat making technologies in music therapy with young people," *Music Therapy Perspectives*, vol. 37, no. 1, pp. 55–64, 2019. https://doi.org/10.1093/mtp/miy025

[6] G. Chen, "Design and application of scenario-based perception of smart wearable device interaction method," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 1, pp. 69–81, 2024. https://doi.org/10.3991/ijim.v18i13.49071

[7] C. E. Burns, "Therapeutic uses of the flute within music therapy practice," *Music Therapy Perspectives*, vol. 37, no. 2, pp. 169–175, 2019. https://doi.org/10.1093/mtp/miz003

[8] M. J. Silverman and J. Leonard, "Effects of active music therapy interventions on attendance in people with severe mental illnesses: Two pilot studies," *The Arts in Psychotherapy*, vol. 39, no. 5, pp. 390–396, 2012. https://doi.org/10.1016/j.aip.2012.06.005

[9] A. Ramachandran, C. Sarabu, U. Gupta, S. Ghose, and V. S. Lee, "Sustainably advancing health AI: A decision framework to mitigate the energy, emissions, and cost of AI implementation," *NEJM Catalyst Innovations in Care Delivery*, vol. 6, no. 10, 2025. https://doi.org/10.1056/CAT.25.0125

[10] G. V. Singh, M. Firdaus, D. S. Chauhan, A. Ekbal, and P. Bhattacharyya, "Zero-shot multitask intent and emotion prediction from multimodal data: A benchmark study," *Neurocomputing*, vol. 569, p. 127128, 2024. https://doi.org/10.1016/j.neucom.2023.127128

[11] C. Curis, "Home-based music therapy for patients with chronic disease through digital technology," *Information and Communication Technology in Musical Field*, vol. 9, no. 2, pp. 57–61, 2018.

[12] L. Văduva and C. Warner, "A case study on music therapy sessions with a bereaved child, and the use of digital devices," *ICT in Muzical Field/Tehnologii Informatice si de Comunicatie in Domeniul Muzical*, vol. 13, no. 2, pp. 19–25, 2022. https://doi.org/10.47809/ICTMF.2022.01.02

[13] T. N. Chien *et al.*, "Usability evaluation of mobile medical treatment carts: Another explanation by information engineers," *Journal of Medical Systems*, vol. 36, no. 3, pp. 1327–1334, 2012. https://doi.org/10.1007/s10916-010-9593-x

[14] P. Marques, P. Váz, J. Silva, P. Martins, and M. Abbasi, "Real-time gesture-based hand landmark detection for optimized mobile photo capture and synchronization," *Electronics*, vol. 14, no. 4, p. 704, 2025. https://doi.org/10.3390/electronics14040704

[15] L. Shen *et al.*, "A first look at generative artificial intelligence based music therapy for mental disorders," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 3, pp. 7439–7453, 2024. https://doi.org/10.1109/TCE.2024.3514633

[16] Y. Li, X. Li, Z. Lou, and C. Chen, "Long short-term memory-based music analysis system for music therapy," *Frontiers in Psychology*, vol. 13, p. 928048, 2022. https://doi.org/10.3389/fpsyg.2022.928048

7 AUTHOR

Ruqi Bai studied at Jiangnan University and obtained her bachelor's degree in 2010. Then, from 2010 to 2013, she pursued her master's degree at Nanjing Normal University and was awarded the master's degree in 2013. In December 2013, she joined XinZhou Normal University. From 2018 to 2019, she conducted exchange studies as a senior visiting scholar at Capital Normal University. She has presided over two provincial-level projects, one horizontal project, and one college student innovation and entrepreneurship project, and has published several academic papers (E-mail: 18603502125@163.com).