Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39095

Artificial Intelligence Feedback Based Investigation of Emotional Regulation Effects in Music Learning

Binru Zou

Conservatory of Music, Shanghai University, Shanghai, China
wellington589125@gmail.com

Abstract. With increasing use of artificial intelligence for performance evaluation in music education, repeated mistakes may cause frustration, anxiety, and even withdrawal from practice due to lack of proper feedback mechanisms addressing such problems. This study developed a multimodal feedback system based on MIDI performance, facial expression coefficients, and interaction logs where an attention-based BiLSTM predicted valence, arousal, and need for high regulation and then used contextual Thompson sampling to select affect-adaptive feedback. In the pilot study, the dataset consisted of 1,080 practice segments from 45 novice piano players, and the six-week experiment included 90 subjects allocated into affect-adaptive, fixed technical, and metric-only feedback conditions. The full multimodal model obtained the following scores: Macro-F1 = 0.873, AUC = 0.936, ECE = 0.029, and MEEL = 46.7 ms. At posttest, note accuracy for the adaptive condition amounted to 93.1%, while in two other conditions, it amounted to 89.4% and 87.5%. Also, restart latency for the adaptive condition reduced to 7.3 seconds. Moreover, multilevel mediation analysis revealed a significant indirect effect through emotional recovery, $\beta = 0.16$, 95% CI [0.07, 0.27]. This research shows that integrating affect recognition with feedback selection is a promising direction for improving computational responsiveness and emotional regulation in music learning. Proposed architecture maintains boundaries set by teacher-reviewed feedback and separates performance and outcome evaluations.

Keywords: artificial intelligence feedback, music learning, emotion regulation, multimodal affective computing, adaptive learning

1. Introduction

Music education is undergoing a significant transformation due to automated performance evaluation, tailored practice, and feedback provision by artificial intelligence (AI). Recent computational surveys indicate the increased importance of deep learning, transformers, and generation approaches in music analysis and interactive assistance [1]. The music-education literature stresses the necessity of hybrid AI-human teaching for combining personalization and pedagogy [2]. Currently, there are no existing systems that focus not only on pitch, rhythm, and notation accuracy but also on the change of attention, perseverance, and motor coordination resulting from the negative experience of repeating errors. This change can be detected by affective computing and then converted into the corresponding feedback using an adaptive decision model.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39095

The proposed research combines three concepts into one closed loop, namely performance assessment, emotion estimation, and sequential feedback selection, and investigates whether emotion-aware adaptation can positively influence emotional recovery, continuous practice, note accuracy, rhythmic consistency, memory, and transfer of knowledge to new repertoire during a six-week controlled piano-playing experiment.

2. Literature review

2.1. Automated performance assessment

Automatic music performance assessment is becoming more likely to model piano playing as structured time series instead of one overall score. Researchers used MIDI events to correlate with the score and demonstrated the effectiveness of performance features in hierarchy form to enhance perception-based evaluations [3]. Some previous piano teaching studies have implemented bidirectional LSTM networks in evaluating MIDI time sequences to lower subjectivity in performance evaluation [4].

2.2. Multimodal emotion recognition

Multimodal emotion recognition in education increasingly uses both behavioral and computational cues due to the complexity of inferring affective states from only one of them. Researchers concluded that datasets that were multimodal and self-built become particularly important if an emotion recognition is implemented in real learning environments [5]. The research on multimodal emotions highlights signal alignment, complementary representation, and fusion as key factors determining the quality of the process [6].

2.3. Adaptive feedback strategy

Adaptive feedback entails the need for a decision-making approach that will adjust to changes in the learner's state and not reapply the same rule again and again. Digital adaptive feedback has the capability of influencing self-regulated learning if the support is customized to each learning style [7]. Recommendations driven by explainable AI have indicated that feedback that is based on data has the ability to assist in monitoring and making adjustments behaviorally during the learning process [8]. Recent studies have also explored the use of large language models to generate in-context, personalized feedback for intelligent tutoring systems, providing a basis for tailoring feedback to learners' specific contexts [9].

3. Experimental methods

3.1. Multimodal dataset construction

The multimodal platform comprised a React practice interface linked with a FastAPI backend via WebSocket to ensure synchronous data transmission with minimal latency. Altogether, forty-five beginning pianists performed 24 pilot practice sessions, generating 1,080 labeled segments for training models, while another group consisting of 90 participants was used in the experiment. The practice platform featured an 88-key MIDI digital piano that logged notes together with timestamps and velocities, while the facial behavior of participants was recorded by a standard 1080p camera in front view. Pitch accuracy, onset offset, duration ratio, tempo inconsistency, and velocity data were

DOI: 10.54254/2755-2721/2026.39095

extracted for each performance window using the pretty_midi library. Facial video streams were transformed into numerical representation of facial expressions using MediaPipe, and the backend collected additional data such as pause duration, number of retries, feedback review time, and time to restart. Synchronization was done by participant ID, session ID, and timestamp of backend log in 30-second windows with z-standardization based on training folds. After each practice session, learners rated their experience with the Self-Assessment Manikin and a frustration item on a five-point scale. Missing windows due to tracking problems were not interpolated but marked.

3.2. Emotion recognition model implementation

MIDI sequences were encoded with a one-dimensional convolutional network using 64 and 128 filters with kernel size 3, while facial coefficients and interaction logs were processed by separate two-layer multilayer perceptrons. Their 128-dimensional embeddings were concatenated and passed to a 128-unit bidirectional LSTM because emotional regulation develops across consecutive practice moments and depends on both preceding errors and subsequent behavioral recovery [10]. Temporal importance was estimated using an attention mechanism, as shown in Equation (1).

$$\alpha_t=\frac{\exp\left(v^T\tanh\left(W_h h_t+b_h\right)\right)}{\sum_{j=1}^{T}\exp\left(v^T\tanh\left(W_h h_j+b_h\right)\right)},$$

$$c=\sum_{t=1}^{T}\alpha_t h_t \tag{1}$$

Where h_t denotes the BiLSTM hidden state at time step t, α_t is its attention weight, and c represents the fused temporal feature. The network jointly predicted valence, arousal, and high regulation need using the multitask loss function shown in Equation (2).

$$\mathscr{L}=\lambda_v MSE(\widehat{V},V)+\lambda_a MSE(\widehat{A},A)+\lambda_r BCE(\widehat{R},R) \tag{2}$$

Where V, A, and R denote the ground-truth valence, arousal, and regulation-need labels, respectively, while λ_v, λ_a, and λ_r control task-specific loss weights. Adam optimization used a learning rate of 0.001, batch size 32, dropout of 0.30, and early stopping. Participant-grouped five-fold cross-validation compared XGBoost, unimodal networks, and attention ablations, followed by calibration and inference-latency analysis.

3.3. Adaptive feedback experiment

After model selection, network weights were frozen and predicted affective states entered a contextual Thompson-sampling feedback policy. The context vector contained regulation probability, dominant performance error, previous-block improvement, cumulative practice duration, and prior feedback response. A standardized reward integrated musical improvement and emotional regulation, as shown in Equation (3).

$$r_t=0.40\Delta z(P_t)+0.25\Delta z(V_t)+0.20\Delta z(-F_t)+0.10\Delta z(-L_t)+0.05C_t \tag{3}$$

DOI: 10.54254/2755-2721/2026.39095

Where P_t denotes performance improvement, V_t represents valence recovery, F_t is frustration intensity, L_t denotes restart latency, and C_t indicates practice-block completion [11]. Thompson sampling selected feedback according to posterior uncertainty, as shown in Equation (4).

$$\tilde{\theta}_a \sim N\left(\hat{\theta}_a, v^2 B_a^{-1}\right),$$

$$a_t = \arg\max_{a \in A}\left(\mathbf{x}_t^T \tilde{\theta}_a\right) \tag{4}$$

Where x_t represents the current learner context, $\tilde{\theta}_a$ is a posterior sample of the parameter vector associated with feedback action a, and a_t denotes the selected teacher-reviewed template. A total of 90 participants was randomly assigned to equal numbers in three conditions, including affect adaptive, fixed technical, and metric only. The training procedure took six weeks, during which two 30-min sessions per week were performed. Each training session consisted of a warm up, practice, state measure, feedback, and replay phases. Assessments were conducted at baseline, week 3, post test, and a two-week follow-up stage, involving trained and untrained music pieces.

4. Results

4.1. Computational model performance

The validation using five-fold participant groupings yielded notable improvements due to multimodal temporal fusion. As presented in Table 1, the attention-BiLSTM for the entire architecture achieved the Macro-F1 score of 0.873 ± 0.019 and AUC of 0.936 ± 0.012, with valence and arousal having their mean absolute errors decrease to 0.512 ± 0.028 and 0.487 ± 0.025, respectively. Expected calibration error is 0.029 ± 0.006, reflecting the alignment between the estimated confidence and the actual results. The architecture without attention achieved the Macro-F1 score of 0.821 ± 0.024. Omitting interaction logs dropped the Macro-F1 to 0.826 ± 0.023 and caused an increase in valence MAE to 0.604 ± 0.031, implying that the interaction data contributed additional emotional information aside from MIDI and facial input data. The mean latency time is 46.7 ± 5.8 ms with the 95th percentile being 58.9 ms.

Table 1. Computational performance of emotion recognition models

Model	Valence MAE	Arousal MAE	Valence CCC	Arousal CCC	Macro-F1	AUC	ECE	Latency (ms)
MIDI only	0.718 ± 0.041	0.681 ± 0.038	0.623	0.598	0.742 ± 0.031	0.812 ± 0.024	0.071 ± 0.009	24.3 ± 3.1
Facial only	0.694 ± 0.037	0.653 ± 0.034	0.647	0.621	0.758 ± 0.028	0.828 ± 0.021	0.066 ± 0.008	31.6 ± 4.2
Interaction only	0.761 ± 0.045	0.733 ± 0.042	0.581	0.552	0.721 ± 0.034	0.791 ± 0.026	0.079 ± 0.010	18.5 ± 2.7
Fusion without attention	0.603 ± 0.033	0.579 ± 0.031	0.741	0.718	0.821 ± 0.024	0.895 ± 0.016	0.043 ± 0.007	39.8 ± 5.1

DOI: 10.54254/2755-2721/2026.39095

Table 1. (continued)

Full model without interaction logs	0.604 ± 0.031	0.568 ± 0.030	0.756	0.734	0.826 ± 0.023	0.901 ± 0.015	0.041 ± 0.007	42.1 ± 5.4
Attention-BiLSTM multimodal model	0.512 ± 0.028	0.487 ± 0.025	0.824	0.801	0.873 ± 0.019	0.936 ± 0.012	0.029 ± 0.006	46.7 ± 5.8

4.2. Learning regulation effects

All 90 participants completed the six-week intervention. As shown in Figure 1, baseline note accuracy was comparable across the affect-adaptive, fixed-technical, and metric-only groups at 82.6%, 82.4%, and 82.5%, respectively. By posttest, accuracy increased to 93.1 ± 4.3%, 89.4 ± 5.0%, and 87.5 ± 5.5%, producing a significant group-by-time interaction, $F(6,261)=8.72$, $p<0.001$, partial $\eta^2=0.167$. The adaptive group retained 92.4 ± 4.5% accuracy at follow-up, whereas the fixed and metric groups declined to 88.8 ± 5.2% and 86.9 ± 5.6%. Mean restart latency in the adaptive group fell from 14.7 ± 4.1 s to 7.3 ± 2.6 s, compared with 10.9 ± 3.1 s and 12.2 ± 3.5 s at posttest. Unfamiliar-piece ratings reached 4.31 ± 0.39, 3.87 ± 0.44, and 3.62 ± 0.47 out of five. Multilevel mediation identified a significant indirect effect through emotional recovery, $\beta=0.16$, 95% CI [0.07, 0.27].

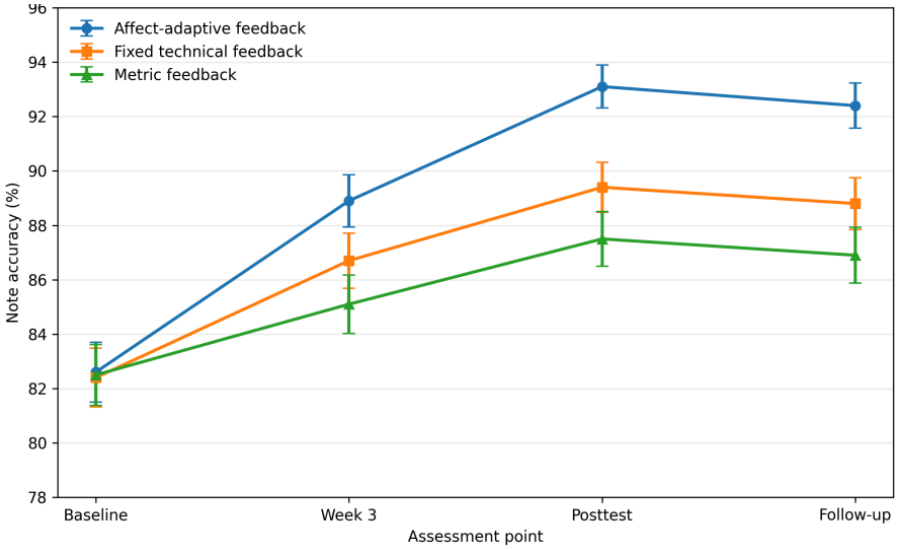


Figure 1. Note accuracy across four assessment points

5. Discussion

The findings seem to imply that feedback becomes an intervention based on a particular emotional state, rather than being a report on a mistake. The biggest improvement in computation seems to happen in the case where interaction logs are maintained, which implies that frustration in practice may be shown by behavior tempo, and not facial expressions or mistakes. A student will remain fairly calm emotionally with facial expressions, but show signs of hesitation, reiteration and delays in response. The post-test and follow-up differences imply that adaptive feedback makes it possible to continue training after mistakes, while technical correction comes after recovery of attentional control. This is supported by the mediation pathway, though emotion regulation cannot be

DOI: 10.54254/2755-2721/2026.39095

considered as the only factor in this process. Attention personalization, novelty and responsiveness of the system may play a role as well. The context-based exploration brings another compromise into the picture, since exploration of the correct feedback policy requires a trial and error approach, but can decrease pedagogical value of the training temporarily.

6. Conclusion

The multimodal feedback model using the combination of MIDI information, face behavior, interaction records, attention-based BiLSTM state estimation, and context-aware Thompson sampling was examined for piano training. The model showed efficient estimation of regulation needs (Macro-F1 = 0.873) and facilitated more rapid restarting behavior, increased posttest accuracy of notes, and increased follow-up retention in the affect-adaptive group. Recovery of emotional state partially acted as an intermediary variable between the computational estimation of a state and the learning behavior. The drawbacks of this research include the usage of the sample of novices, a relatively short period of six weeks, dependence on the video camera to obtain the facial features, and limited feedback space.

References

[1] Mycka, J., & Mańdziuk, J. (2025). Artificial intelligence in music: Recent trends and challenges. *Neural Computing and Applications, 37*(2), 801–839.
[2] Mazlan, C. A. N., et al. (2026). Artificial intelligence applications and pedagogical challenges in music education. *Discover Education, 5*(1), 140.
[3] Park, J., et al. (2024). Piano performance evaluation dataset with multilevel perceptual features. *Scientific Reports, 14*(1), 23002.
[4] Wei, X., & Sun, S. (2024). Optimization of piano performance teaching mode using network big data analysis technology. *International Journal of Information and Communication Technology Education (IJICTE), 20*(1), 1–20.
[5] Yu, S., et al. (2024). Bridging computer and education sciences: A systematic review of automated emotion recognition in online learning environments. *Computers & Education, 220*, 105111.
[6] 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.
[7] Mejeh, M., Sarbach, L., & Hascher, T. (2024). Effects of adaptive feedback through a digital tool–a mixed-methods study on the course of self-regulated learning. *Education and Information Technologies, 29*(14), 1–43.
[8] Afzaal, M., et al. (2024). Informative feedback and explainable AI-based recommendations to support students' self-regulation: M. Afzaal et al. *Technology, Knowledge and Learning, 29*(1), 331–354.
[9] Reddig, J. M., Arora, A., & MacLellan, C. J. (2025). Generating in-context, personalized feedback for intelligent tutors with large language models. *International Journal of Artificial Intelligence in Education, 35*(6), 3459–3500.
[10] Wang, S. (2025). Hybrid models of piano instruction: How combining traditional teaching methods with personalized AI feedback affects learners' skill acquisition, self-efficacy, and academic locus of control. *Education and Information Technologies, 30*(9), 12967–12989.
[11] De Kerpel, L., Thuy, A., & Benoit, D. F. (2026). A bandit-based approach to educational recommender systems: Contextual Thompson sampling for learner skill gain optimization. *INFORMS Transactions on Education, 26*(3), 187–199.