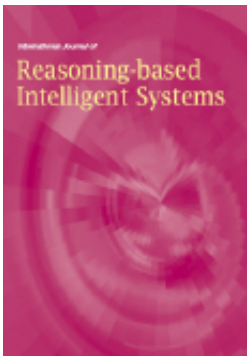
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

A study on piano performance training and classroom feedback methods in AI-assisted music performance instruction

Hua Wang

DOI: 10.1504/IJRIS.2026.10081485

Article History:

Received:	30 June 2026
Last revised:	13 August 2026
Accepted:	13 August 2026
Published online:	20 October 2026

Copyright © 2026 Inderscience Enterprises Ltd.

A study on piano performance training and classroom feedback methods in AI-assisted music performance instruction

Hua Wang

School of Music,
Nangjing University of the Arts,
Nanjing, 210013, Jiangsu, China
Email: wanghua@nua.edu.cn

Abstract: AI-assisted piano instruction requires reliable identification of performance errors and timely conversion of diagnostic evidence into actionable classroom feedback. An event-level multimodal framework is developed using synchronised audio, MIDI, hand-pose, and score information. Its methodological core is a score-conditioned event reliability fusion mechanism, which dynamically weights modality-specific evidence according to score-referenced deviations and branch uncertainty. The model-development dataset contains 156 students from four institutions, with participant-disjoint training and testing. The proposed model achieves an F1-score of 92.2% with an average inference latency of 41.3 ms, while external-institution evaluation retains an F1-score of 90.5%. In the eight-week classroom experiment, rhythm deviation decreases from 79.1 ms to 42.7 ms, note omission rate decreases from 6.9% to 3.1%, and instructor score increases to 86.9 points. The framework is intended for technical performance diagnosis and formative feedback, covering pitch, timing, fingering, dynamics, pedal use, legato continuity, while higher-level artistic interpretation remains under instructor judgment.

Keywords: AI-assisted instruction; piano performance training; multimodal fusion; error detection.

Reference to this paper should be made as follows: Wang, H. (2026) 'A study on piano performance training and classroom feedback methods in AI-assisted music performance instruction', *Int. J. Reasoning-based Intelligent Systems*, Vol. 18, No. 28, pp.87–99.

Biographical notes: Hua Wang received her Master of Piano Performance, PhD in Education, Associate Professor and Master's Supervisor at the School of Music, Nanjing University of the Arts. She has won multiple government-level awards in piano performance. Her research focuses on the integration of music practice and theory. She has presided over and participated in provincial/ministerial and national-level projects, and many papers have been published in national and provincial core journals. In teaching, she empowers students through a professionalised teaching system, leading students to numerous awards in national and provincial piano competitions and cultivating many outstanding graduates.

1 Introduction

Artificial intelligence is increasingly used to convert complex musical performance processes into measurable representations for transcription, performer analysis, movement interpretation, and technology-assisted instruction. Recent work has extended this direction from single-channel acoustic analysis toward multimodal and behaviour-aware modelling. Younesi Heravi et al. (2026) used vision-derived kinematic features for human activity intensity estimation, illustrating the feasibility of quantifying motion from structured visual descriptors. In piano-related analysis, Mycka and Mańdziuk (2026) applied deep learning to piano-performer classification, showing the growing role of learned representations in distinguishing performance characteristics. Emerson and Harrison (2025) further organised multimodal datasets for studying expert performances of musical scores, highlighting the importance of jointly considering complementary performance information rather than relying on audio alone. Kappan et al. (2025) reviewed deep-learning approaches for two-and three-dimensional human pose estimation, providing a methodological basis for extracting finger, wrist, and hand-motion descriptors from video. In instructional settings, Deja et al. (2025) developed an augmented-reality piano-roll system for improvisation teaching, indicating that interactive technologies can participate directly in piano learning rather than serving only as post-performance analysis tools. Patil et al. (2025) summarised machine-learning methods, datasets, and challenges for music-emotion analysis, demonstrating that musical performance assessment also contains expressive dimensions extending beyond pitch and timing accuracy. Kim et al. (2025) proposed a periodicity-frequency fusion network for multi-

Copyright © The Author(s) 2026. Published by Inderscience Publishers Ltd. This is an Open Access Article distributed under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

instrument music transcription, reflecting continued development of learned representations for note-event analysis. At a broader level, Ghosh et al. (2025) discussed risks and governance frameworks associated with artificial intelligence, reinforcing the need to define automated assessment boundaries and preserve instructor oversight when model outputs are used educationally. Dibenedetto et al. (2025) compared deep-learning approaches to human pose estimation, further supporting the use of vision-based movement representations. Mycka and Mańdziuk (2025) reviewed recent trends and challenges in artificial intelligence for music, situating these technical developments within the broader evolution of computational music analysis.

These studies establish useful foundations in music transcription, performer modelling, pose estimation, multimodal data organisation, and technology-supported music learning, but they leave a specific gap in piano classrooms. Existing approaches typically quantify acoustic output, classify performers, estimate body motion, or provide interactive teaching functions as separate tasks. A classroom diagnosis system must additionally determine which note event is problematic, identify whether the evidence originates from sound, keystroke timing, or hand motion, and convert that evidence into feedback that can be acted on during practice. Direct concatenation of audio, MIDI, and pose features is insufficient when one modality is unreliable at a particular note event, while unconstrained attention may assign large weights to latent features that are inconsistent with score-defined pitch, duration, dynamics, or fingering requirements. To address this problem, an event-level multimodal framework is developed using MIDI-indexed note events as shared computational anchors. Audio, MIDI, and hand-pose branches encode spectral output, keystroke behaviour, and movement processes, respectively, while score-derived references provide explicit constraints for cross-modal comparison. The methodological contribution lies in a score-conditioned event reliability fusion mechanism rather than in introducing new CNN, BiLSTM, GCN, or DTW operators. For each note event, modality-specific deviations from score references are combined with branch uncertainty to estimate reliability, and these reliability terms regulate the contribution of each branch during fusion. The fused representation is then used for technical-error recognition and is mapped to prioritised classroom feedback and personalised training routes, forming a diagnosis-to-intervention pathway. The intended application boundary is technical performance diagnosis: the automated system evaluates pitch, rhythm, fingering, dynamics, pedal use, and legato continuity, while phrasing, tone shaping, structural awareness, stylistic interpretation, and emotional expression remain under instructor judgment. Evaluation therefore focuses on the accuracy, robustness, generalisability, computational efficiency, and instructional utility of technical-error recognition rather than treating technical accuracy as a complete measure of piano artistry. Under participant-independent testing, the complete model achieved an F1-score of 92.2%, while classroom evaluation examined technical improvement, feedback efficiency, and individualised practice outcomes, providing evidence for the framework within its explicitly defined instructional scope.

Table 1 Study participants and experimental data composition

Item	*Data composition*	*Parameter or quantity*	*Purpose*
Research subjects	Beginner, intermediate, and advanced students	156 students from four institutions	Training results and model generalisation validation
Performance materials	Scales, arpeggios, etudes, counterpoint, and musical excerpts	62 sets	Covering different technical tasks
Audio data	Mono waveform files	44.1 kHz, 24-bit	Extract spectral, energy, and timbre features
MIDI data	Key-on, key-off, velocity, and pedal events	Approximately 186,000 entries	Create event-level performance sequences
Video data	Hand motion sequences	1080 p, 60 fps	Extraction of finger and wrist pose features
Valid clips	Cleaned performance clips	14,040 segments	Model training and testing
Error labels	8 error classes	112,320 event labels	Error identification supervised learning
Data split	Participant-disjoint training, validation, and internal test sets	82/23/12 students; five-fold subject-independent cross-validation on 117 development participants; external test cohort: 39 students	Model training and performance evaluation

2 Materials and methods

2.1 Study participants and experimental materials

The model-development dataset included 156 students recruited from four independent music-education institutions, all with continuous piano training experience. They were classified as beginner, intermediate, and

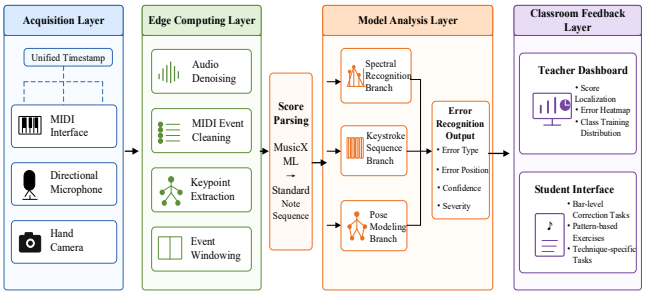
advanced according to entrance performance tests and instructor evaluation, with 52 students at each proficiency level. The experimental materials were expanded to 62 sets of scales, arpeggios, etudes, contrapuntal passages, and excerpts covering classical, romantic, and twentieth-century repertoire, thereby increasing variation in tempo, articulation, texture, fingering, dynamics, and pedal usage. Data from three institutions were used for model development, whereas all data from the fourth institution were reserved as an independent external test cohort. Using a digital piano, audio recording equipment and video recording, performance data was simultaneously recorded. Key numbers, key press time, key release time, velocity values and pedal status were recorded using the musical instrument digital interface (MIDI), audio signals at 44.1 kHz and 24-bit resolution, and hand movements with a 1080p, 60 fps camera. The location in the sheet music was used for annotating the error category, error categories were reviewed by two piano instructors, and the error categories were entered to the sample database. The composition of the data is shown in Table 1.

2.2 *Architecture of the AI-assisted piano training system*

As shown in Figure 1, the system uses a collaborative architecture involving data acquisition on device, edge inference and feedback in classroom. The acquisition end includes a MIDI interface, a directional microphone and a hand tracking camera, each type of data is synchronised by a unified clock and sent to the edge computing unit. The edge unit contains a lightweight graphics processing unit (GPU) along with 16 GB of memory and an Intel i7 processor, which are used for the audio denoising, MIDI event cleaning, video keypoint extraction and event window segmentation. The model layer includes three branches: spectral recognition, keystroke sequence, and posture modelling, to produce the output: error category, error location, confidence and severity (Volta and Di Stefano, 2024). At the classroom end there is a teacher visualisation interface and student training interface. Sheet music positioning, error hot spots, and class training distribution are provided on the teacher interface and error-correction tasks based on measure, musical phrase and technique type are delivered to the student interface.

The system can be divided into three entities that can be computed: musical score events, audio output and motion processes, as illustrated in Figure 1. The musical score parsing module parses a MusicXML file to generate standard note sequences; audio frames are used to validate the audio output; and video keypoints can be used to find discrepancies in hand shape and fingering (Burunat et al., 2024). A model analysis layer creates multimodal feature representations for each note event, and the error identification results are saved back to the score coordinates, along with the timestamps. The classroom feedback layer sends prompts according to the severity of the errors and the importance of instruction, which will not only give the overall feedback score evaluation but also help the teacher to locate the specific measures, voices and technical actions.

Figure 1 Architecture of the AI-assisted piano training system (see online version for colours)



2.3 *Multimodal performance data acquisition*

For data acquisition, all the audio and video frames are mapped to the same note event window, according to the MIDI key-press event sequences. The processing on the audio side consists of pre-weighting, short-time framing and spectral subtraction noise reduction with a 2,048 sample frame length and a 512 sample frame shift; on the video side, the coordinates of 21 hand keypoints are extracted from each video frame, and the keypoints with confidence below 0.60 are discarded; on the MIDI side, erroneous touch events that do not span for more than 30 ms are removed (Christodoulou et al., 2024). Since sampling frequencies of different modalities are inconsistent, time alignment is performed using equation (1):

$$k_i^A = \arg\min_k \left| \frac{kH}{f_s} - t_i^M \right|, \quad k_i^V = \arg\min_k \left| \frac{k}{f_v} - t_i^M \right| \tag{1}$$

where k_i^A denotes the audio frame number corresponding to the i^{th} MIDI event, k_i^V denotes the corresponding video frame number, H denotes the audio frame offset, f_s denotes the audio sampling rate, f_v denotes the video frame rate, and t_i^M denotes the key press time of the i^{th} MIDI event. Equation (1) is used to select the closest audio and video frames near each key-on point and to construct an event window using $\left[t_i^M - 120ms, t_i^M + 300ms\right]$, ensuring that the preparatory movement before key-on, the moment of key-on, and the sound decay after key-off are all included in the sample (Cho et al., 2024).

2.4 *Piano performance feature extraction*

A score-aligned note event is defined as the basic unit for performance feature construction, as they are used to build the performance features. On the audio side, the following features are extracted: constant-Q transform (CQT) spectrogram, MEL-frequency cepstral coefficients (MFCC), short-term energy, spectral centre of mass, and harmonic energy ratio; on the MIDI side, are extracted key-on time, duration, velocity, interval between adjacent

notes, pedal overlap rate, and event density for the left and right hands; on the posture side, we extract fingertip velocity, joint angles, wrist height, palm displacement, and interdigital distance; on the score side, we extract standard pitch, standard duration, dynamics markings, voice number, and suggested fingering. After standardisation, features from each modality are concatenated into event-level vectors:

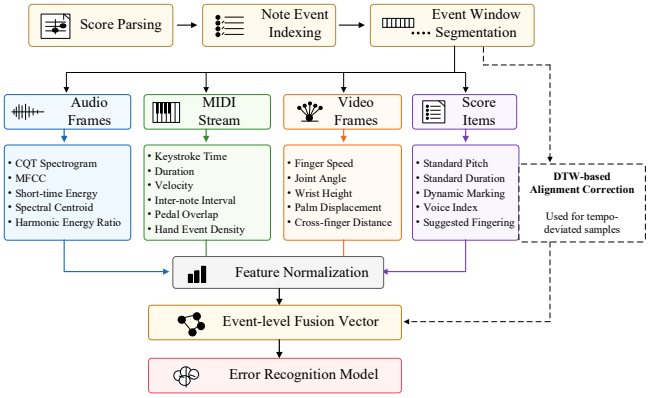
$$\mathbf{x}_i = [\mathbf{a}_i; \mathbf{m}_i; \mathbf{p}_i; \mathbf{s}_i] \tag{2}$$

where $\mathbf{x}_i$ represents the fused feature vector for the i^{th} note event (i), $\mathbf{a}_i$ represents the audio feature vector, $\mathbf{m}_i$ represents the MIDI touch feature vector, $\mathbf{m}_i$ represents the posture feature vector, and $\mathbf{p}_i$ represents the musical notation constraint feature vector (Colafiglio et al., 2024). To mitigate the impact of individual student dynamics and camera mounting angles on the model, features in each dimension are normalised according to equation (3):

$$\tilde{x}_{i,d} = \frac{x_{i,d} - \mu_d}{\sigma_d + \varepsilon} \tag{3}$$

In the equation, $\tilde{x}_{i,d}$ represents the normalised feature in the d^{th} dimension for the i^{th} event; $x_{i,d}$ represents the original feature value; μ_d and σ_d represent the mean and SD, respectively, of the d^{th}-dimensional feature in the training set; and ε is a smoothing term to prevent the denominator from becoming zero, set to 10−6.

Figure 2 Event-level feature construction process for piano performances (see online version for colours)



As shown in Figure 2, the feature construction process extracts audio results, keystroke behaviour, and motion sequences in a single event index, thus eliminating the ambiguity in the positioning of events based on the full audio segment. A technique known as dynamic time warping (DTW) is used inside the event window to compensate for nonlinear distortion of the music sequence from the performance and the standard musical score, so that the tempo changes do not appear as missed or spurious notes (Kyriakou et al., 2024).

2.5 Construction of the performance error recognition model

For error recognition, a multi-branch model is constructed for audio, MIDI, and hand-pose representations. The audio branch uses a three-layer CNN with 32, 64, and 128 channels and kernel sizes of 3 × 3, 3 × 3, and 5 × 5, respectively. Each convolution is followed by batch normalisation, ReLU activation, and 2 × 2 max pooling, and the final feature map is projected to a 128-dimensional event representation. The MIDI branch contains a two-layer bidirectional LSTM with 128 hidden units in each direction, followed by a 256-to-128 linear projection. A dropout rate of 0.30 is applied between recurrent layers. The pose branch constructs a spatial graph from 21 hand keypoints and employs three GCN layers with hidden dimensions of 64, 128, and 128; each layer uses ReLU activation and dropout of 0.25. Graph edges follow the anatomical hand topology, connecting the wrist to metacarpophalangeal joints and adjacent joints sequentially along each finger; self-loops are added and the adjacency matrix is symmetrically normalised before graph convolution. The three branches therefore output representations with the same dimension before reliability-based multimodal fusion. After fusion, the 128-dimensional event representation is passed through a 128–64 fully connected layer with ReLU and dropout of 0.30, followed by a 64–8 output layer and softmax classification for the eight error classes (Louro et al., 2024).

The three branch representations are not fused using unconstrained attention alone. A score-conditioned event reliability fusion mechanism is introduced to estimate the reliability of each modality for every note event. For modality m, the deviation between the branch observation and its score-derived reference is first calculated, and the deviation is combined with branch prediction uncertainty to obtain an event-specific reliability coefficient:

$$\mathbf{d}_i^m = \left\| \mathbf{P}_m \mathbf{h}_i^m - \mathbf{s}_i^m \right\|_1, \; r_i^m = \frac{\exp\left(-d_i^m\right)}{u_i^m + \varepsilon} \tag{4}$$

The reliability coefficient is then introduced into the attention score so that modalities consistent with the score reference and having lower prediction uncertainty receive larger weights:

$$\eta_i^m = \frac{\exp\left[g_m\left(\mathbf{h}_i^m\right) + \lambda \ln\left(r_i^m + \varepsilon\right)\right]}{\sum_{m \in \{A,M,P\}} \exp\left[gm(hin) + \lambda ln(rin + \varepsilon)\right]}, \mathbf{h}_i = \sum_{m \in \{A,M,P\}} \eta_i^m \mathbf{h}_i^m \tag{5}$$

For each modality, the scoring function is implemented as a two-layer multilayer perceptron with dimensions 128–64–1, using a tanh activation in the hidden layer. The score-derived reference projection P_m maps each branch representation from 128 dimensions to a 32-dimensional constraint space. The reliability-strength coefficient λ is set to 0.40 based on the validation set. This mechanism makes the fusion process dependent on explicit performance evidence rather than only latent feature similarity. For example, MIDI timing evidence receives a larger

contribution for rhythm errors, whereas pose reliability increases for fingering-path and wrist-motion errors. A multi-task loss function is adopted during the training phase:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{pitch} + \lambda_2 \mathcal{L}_{rhythm} + \lambda_3 \mathcal{L}_{velocity} + \lambda_4 \mathcal{L}_{gesture} \quad (6)$$

where $\mathcal{L}$ denotes the total loss, $\mathcal{L}_{pitch}$ denotes the pitch error loss, $\mathcal{L}_{rhythm}$ denotes the rhythm error loss, $\mathcal{L}_{velocity}$ denotes the dynamics control loss, $\mathcal{L}_{gesture}$ denotes the posture deviation loss, and λ_1 through λ_4 are the task weights. During training, the initial weights are set to 0.30, 0.30, 0.20, and 0.20, respectively. These values were selected on the validation set using a constrained grid search with a step size of 0.05 and $\sum_{k=1}^{4} \lambda_k = 1$. Candidate configurations were ranked by macro-F1 while requiring that recall for any individual error task did not decrease by more than 1.5 percentage points. The optimum was (λ_1, λ_2, λ_3, λ_4) = (0.30, 0.30, 0.20, 0.20). A ±20% perturbation of each weight followed by renormalisation produced macro-F1 values of 91.7%–92.2%, with the maximum task-level recall variation limited to 0.9 percentage points, indicating low sensitivity to moderate weight changes. After the F1-score on the validation set stabilises, fine-tuning is performed using uncertainty-weighted adjustment to prevent high-noise modalities from dominating the classification results (Sujeesha and Mala, 2024). The AdamW optimiser is used with an initial learning rate of 1×10^{-3}, a batch size of 32, and a maximum of 120 training epochs. Weight decay is set to 1×10^{-4}, and the gradient norm is clipped at 5.0. A cosine-annealing learning-rate schedule decreases the learning rate to 1×10^{-5} during training. Early stopping is activated when validation loss does not decrease for 12 consecutive epochs. Convolutional, graph-convolutional, and linear layers use Kaiming initialisation, whereas recurrent weight matrices use Xavier initialisation (Woo et al., 2023).

All experiments were implemented in Python 3.11 using PyTorch 2.4.1 and CUDA 12.1. Audio processing used librosa 0.10.2, hand-keypoint extraction used MediaPipe 0.10.14, and statistical evaluation used scikit-learn 1.5.0. Training was conducted on an NVIDIA RTX 4070 GPU with 12 GB memory. Identical pre-processing, architecture, and optimisation settings were maintained across baseline, ablation, and repeated-seed experiments.

2.6 Classroom feedback generation method

The classroom feedback module inputs the output from the error identification model and transforms the outcomes of error identification to classroom feedback which is interpretable by teachers and training tasks that can be performed by students (Ramoneda et al., 2024). The system sets up a feedback unit for every error type; the content of each unit consists of the error location, the error type, the level of the error, its frequency of occurrence, the confidence level for the error type, the technical points related to the error and the suggested practice segments for the error type. To prevent feedback overload, a priority ranking mechanism is used to filter out errors that require immediate attention in the classroom:

$$R_j = \alpha E_j + \beta F_j + \gamma C_j - \delta I_j \quad (7)$$

where R_j represents the feedback priority for error class j; E_j represents the error severity, weighted by rhythm deviation, pitch deviation, and dynamics deviation; F_j represents occurrence frequency, C_j represents model recognition confidence; and I_j represents the historical improvement rate. The four terms have distinct instructional interpretations: severity reflects the immediate consequence of an error, frequency represents its persistence, confidence suppresses feedback generated from uncertain recognition, and improvement rate reduces repeated prompts for errors already showing stable correction. Rather than being manually fixed, α, β, γ, and δ were selected using an independent pilot dataset containing eight lessons from 12 students who were not included in the 48-student classroom experiment. The coefficients were constrained to be nonnegative and to sum to 1.0. A grid search with a step size of 0.05 evaluated candidate combinations using teacher agreement with the top-three feedback items, first-attempt correction rate, and excessive-prompt rate. The highest validation utility was obtained at α = 0.35, β = 0.25, γ = 0.25, and δ = 0.15; therefore, the numerical values used in the original priority equation were retained, but are now empirically determined rather than heuristically assigned (Edwards et al., 2023). Equation (7) is used to provide early feedback on errors that are high-frequency, high-severity, and have high confidence, while reducing the intensity of prompts for errors that have already shown significant improvement.

The feedback generation process consists of event aggregation, score rewriting and training task matching. The event aggregation stage aggregates up to three consecutive adjacent notes of the same type of error to phrase-level errors; the score rewriting stage maps the error markers to the measures, parts and hands of the score; the training task matching stage extracts similar musical phrase fragments from the exercise library and creates training tasks according to the four training methods: slow hand-separated practice, metronome synchronisation, dynamic grading, and pedal release. Feedback can be edited by the instructor and the modified feedback is added to the feedback sample library, which is used for updates of feedback templates and for incremental model learning.

Weight robustness was further examined by independently perturbing each optimised coefficient by ±20% followed by renormalisation. Sensitivity was assessed using top-three teacher agreement, first-attempt correction rate, and rank consistency of generated feedback priorities. Scenario-specific tests were also performed for rhythm-dominant, technique-dominant, and mixed-repertoire lessons.

2.7 Experimental design and evaluation metrics

The experiment consisted of model performance validation, cross-institution generalisation validation, and classroom application validation. All dataset partitions were performed at the participant level rather than the clip level to prevent performance events from the same student from appearing in both training and testing data. Among the 117 students from the three development institutions, 82, 23, and 12 students were assigned to the training, validation, and internal test sets, respectively. All recordings and event labels generated by one student were assigned exclusively to one subset. To further reduce dependence on a single partition, five-fold subject-independent cross-validation was performed on the 117 development participants using fixed model hyperparameters.

For stochastic models, training was independently repeated five times using random seeds 17, 29, 43, 61, and 79 while maintaining identical subject-level partitions and hyperparameters. Results are reported as mean ± standard deviation (SD). 95% confidence intervals (95% CIs) were estimated by 2,000 participant-level bootstrap resamples. Paired permutation tests were used to compare the proposed model with the strongest baseline using participant-level predictions, and Cohen's d was calculated as the effect size. For classroom outcomes, group × time effects were evaluated by two-way mixed analysis of variance, with partial η_p^2 reported as the effect size (Özer et al., 2023). Statistical significance was defined as $p < 0.05$. Each fold contained 23–24 held-out students, while the remaining participants were used for model fitting. The remaining 39 students from the fourth institution were completely excluded from model development and used only for external validation. Their recordings were collected under the same acquisition protocol but were supervised and annotated by instructors who did not participate in model development. Two instructors from the external institution independently reviewed the error labels without access to model predictions, yielding a Cohen's κ of 0.88. Neither these students nor their instructors, recordings, or annotations contributed to model training, hyperparameter selection, or threshold calibration (Moura et al., 2023). The external model was evaluated directly without fine-tuning. An additional repertoire-holdout test excluded selected musical materials during training to evaluate recognition on unseen pieces.

Performance scores were independently assigned by three piano instructors with 10–18 years of professional teaching experience, none of whom participated in system development (Ostermann et al., 2023). Raters were blinded to the students' group assignment and assessment phase. A 100-point analytic rubric included note and rhythm accuracy (25 points), dynamics and tone control (20 points), articulation and pedal control (15 points), fingering and technical coordination (15 points), phrasing and structural coherence (15 points), and stylistic expression (10 points). The final teacher score was the mean of the three ratings. Inter-rater reliability was high, with an ICC (2, k) of 0.91 (95% CI: 0.87–0.94).

Broader educational outcomes were additionally assessed through behavioural engagement, cognitive workload, usability, satisfaction, instructor acceptance, and delayed retention. Engagement was calculated from assigned-task completion and voluntary repeated-practice records; cognitive workload was assessed using NASA-TLX; students in the AI-assisted group completed the system usability scale (SUS) and a five-point satisfaction questionnaire. A four-week delayed assessment was conducted to evaluate retention after the formal intervention.

The AI-assisted teaching group had the same teaching content as the traditional teaching group, adding error detection, in-class feedback and personalised training path modules (Bontempi et al., 2023) to the traditional teaching group, while the traditional teaching group received only verbal correction by the teacher and recorded after-class exercises. The experimental design and evaluation metrics are shown in Table 2.

Classification performance is calculated using equation (8):

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

where *Precision* denotes the proportion of correctly identified positive events among all events predicted as positive, *Recall* denotes the proportion of correctly identified positive events among all reference positive events, and $F1$ is their harmonic mean. Higher $F1$ values indicate a better balance between recognition precision and detection completeness. Continuous outcomes are evaluated using the root mean square error (RMSE) defined in equation (9):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{9}$$

where y_i is the reference value obtained from the score annotation or instructor assessment, $\hat{y}_i$ is the corresponding system estimate, and n is the number of evaluated samples. Rhythm deviation, dynamics prediction, and feedback delay are all calculated using the same statistical methodology to ensure that model performance metrics can be analysed in correspondence with classroom application metrics.

Table 2 Experimental design and evaluation metrics

Experiment type	*Comparison group*	*Sample size or duration*	*Evaluation criteria*	*Statistical method*
Error detection performance	Threshold rules, SVM, CNN, CNN-BiLSTM, multimodal fusion model	12 participant-held-out students	Accuracy, precision, recall, F1-score, recognition latency	Cross-model comparison
Ablation experiments	Removing audio, removing MIDI, removing pose, eplacing score-conditioned reliability fusion with conventional attention, and replacing attention fusion with equal-weight multimodal fusion	12 participant-held-out students	Accuracy, recall, F1-score, and F1-score change relative to the complete model	Five-seed repeated evaluation and paired permutation testing
Training results	Traditional instruction group, AI-assisted instruction group	8 weeks	Tempo deviation, omission rate, coefficient of variation in dynamics, teacher scores	Between-group and within-group comparisons
Feedback efficiency	Human feedback, AI-assisted feedback	16 class sessions	Error location time, feedback generation time, first-time correction rate	Statistics on mean and standard deviation
Path recommendations	Fixed training routes, personalised training routes	8 weeks	Task completion rate, error reduction rate, learning gain	Phased trend analysis
Generalisation validation	Internal test, external institution, unseen repertoire	12 internal + 39 external students	Accuracy, precision, recall, F1-score	Participant-independent evaluation

3 Results

3.1 Performance results for performance error detection

Threshold rules, support vector machines, convolutional neural networks, and bidirectional long short-term memory networks, and multimodal fusion models have all been verified. Classification precision, recall, accuracy, F1-score, and mean latency for each event were calculated; the results are presented in Table 3.

Table 3 indicates that the multimodal fusion model provides the best classification performance, reaching an accuracy of 94.3±0.3% and an F1-score of 92.2±0.4%. The stronger contemporary baselines substantially narrowed the gap observed with CNN-BiLSTM: the Multimodal Transformer, Graph Transformer, MT2-based model, and MERT-based model achieved F1-scores of 90.9%, 91.1%, 91.3%, and 91.5%, respectively. Nevertheless, the proposed model retained the highest F1-score while requiring only 41.3 ms per event, compared with 71.6 ms for the MERT-based baseline. Compared with the strongest contemporary baseline, MERT, the proposed method improved F1-score by 0.7 percentage points (95% CI: 0.2–1.2, $p = 0.014$, Cohen's $d = 0.68$). The improvement was smaller than that over conventional CNN-BiLSTM, but remained statistically significant across repeated runs. This result suggests that the advantage is attributable to score-conditioned event-level reliability modelling rather than simply using a larger network backbone.

Table 3 Comparison of recognition performance among different models

Model	*Accuracy/ %*	*Precision/ %*	*Recall/ %*	*F1-score/ %*	*Latency/ ms*
Threshold rules	82.6	78.9	74.3	76.5	18.4
SVM	86.1	82.4	80.3	81.3	22.7
CNN	89.7 ± 0.4	87.6 ± 0.5	85.8 ± 0.6	86.7 ± 0.5	27.9±0.4
CNN-BiLSTM	91.8 ± 0.4	89.9 ± 0.5	88.6 ± 0.5	89.2 ± 0.4	34.6±0.5
Multimodal transformer	93.0 ± 0.4	91.4 ± 0.5	90.4 ± 0.5	90.9 ±0.4	52.6±0.9
Graph transformer	93.2 ± 0.4	91.6 ± 0.4	90.6 ± 0.5	91.1 ± 0.4	57.8±1.1
MT2-based model	93.4 ± 0.3	91.8 ± 0.4	90.8 ± 0.4	91.3 ± 0.3	64.2±1.3
MERT-based model	93.6 ± 0.3	92.0 ± 0.4	91.0 ± 0.4	91.5 ± 0.3	71.6±1.5
Proposed model	94.3 ± 0.3	92.8 ± 0.4	91.7 ± 0.4	92.2 ± 0.4	41.3±0.6

Notes: Values for neural models are reported as mean ± SD over five independent runs. Threshold rules and SVM are deterministic under the fixed subject-level partition.

To quantify the contribution of individual modalities and fusion components, controlled ablation experiments were performed by removing one component at a time while

retaining the same subject partition, optimisation parameters, and training procedure. The results are presented in Table 4.

Table 4 Ablation results of multimodal and fusion components

Model configuration	*Accuracy/ %*	*Recall/ %*	*F1-score/ %*	*ΔF1/ pp*	*p-value*
Complete proposed model	94.3±0.3	91.7±0.4	92.2±0.4	–	–
Without audio branch	89.4±0.5	86.2±0.7	86.8±0.6	−5.4	<0.001
Without MIDI branch	90.7±0.5	87.7±0.6	88.4±0.5	−3.8	<0.001
Without pose branch	91.5±0.4	89.0±0.5	89.6±0.5	−2.6	0.002
Conventional attention	92.7±0.4	89.6±0.5	90.3±0.4	−1.9	0.006
Equal-weight fusion	91.9±0.5	89.1±0.6	89.7±0.5	−2.5	0.003

Notes: Values are reported as mean ± SD over five independent runs. ΔF1 denotes the F1-score difference relative to the complete model. p-values were obtained using participant-level paired permutation tests against the complete model.

As shown in Table 4, removing any modality produced a statistically significant performance decrease. Removing the audio branch caused the largest F1-score reduction, from 92.2% to 86.8%, demonstrating that spectral and acoustic information remains the primary evidence for pitch, timbre, and articulation errors. Removing MIDI reduced the F1-score by 3.8 percentage points, confirming the importance of explicit keystroke timing and duration information. Removing pose information reduced the F1-score to 89.6%, indicating that hand-motion evidence provides complementary information for fingering and movement-related errors.Category-level analysis further showed that pose removal produced the largest F1-score decreases for fingering-path deviation (−6.8 percentage points) and legato-break errors associated with unstable finger–wrist motion (−4.3 percentage points), whereas pitch and timing categories changed by no more than 1.2 percentage points.

The architectural ablations further distinguish multimodal composition from the proposed fusion mechanism. Replacing score-conditioned event reliability fusion with conventional attention reduced the F1-score from 92.2% to 90.3%, while replacing attention with equal-weight fusion yielded 89.7%. These results show that the performance gain arises from both complementary modality information and event-specific reliability weighting, rather than merely increasing the number of network branches.

Despite the additional multimodal computation, the complete model maintained an average recognition latency of 41.3 ms, which remains suitable for real-time classroom feedback.

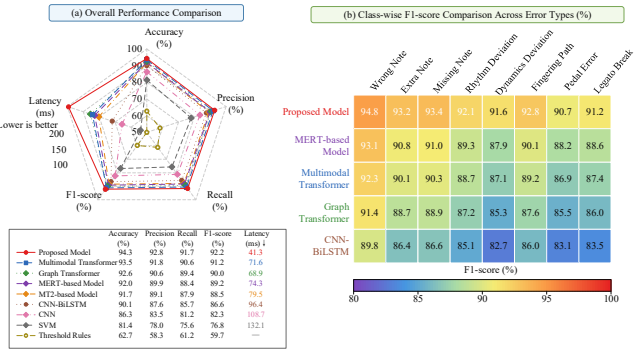
As shown in Table 5, the external-institution cohort achieved an accuracy of 92.5% and an F1-score of 90.5% without model fine-tuning, corresponding to only a 1.7-percentage-point reduction from the internal test. Under unseen-repertoire evaluation, the F1-score remained at 89.8%. Five-fold subject-independent cross-validation further yielded a mean F1-score of 91.8 ± 0.5%, indicating that recognition performance remained stable under participant, institutional, and repertoire shifts.

Table 5 Generalisation performance under different validation conditions

Validation condition	*Accuracy/ %*	*Precision/ %*	*Recall/ %*	*F1-score/%*
Internal participant-held-out test	94.3	92.8	91.7	92.2
External institution	92.5	91.3	89.8	90.5
Unseen repertoire	91.8	90.6	89.1	89.8

To further examine differences among models at the category level, Figure 3 shows the recognition performance for eight error categories. The multimodal fusion model achieved F1-scores of 93.4%, 91.6%, 92.8%, and 90.7% for missed notes, early rhythm, late rhythm, and abnormal dynamics, respectively. Although the threshold-based method retained relatively high performance for pitch and dynamics errors, its performance decreased markedly for fingering-path deviations, pedal blurring, and legato breaks. The multimodal model exhibited less cross-category confusion, indicating that complementary event evidence contributes to more stable error discrimination.

Figure 3 Performance results for performance error detection (see online version for colours)



Computational efficiency was further evaluated to determine whether the proposed model could support real-time classroom deployment. The complete model contains 4.3 million trainable parameters and requires approximately 1.36 GFLOPs for each event-level inference, while the serialised FP32 model occupies 17.4 MB. The 41.3 ms reported in Table 6 represents model inference latency on the RTX 4070 platform; the complete

acquisition-to-feedback pipeline additionally includes multimodal preprocessing, event alignment, and feedback generation. Deployment results on different hardware platforms are shown in Table 6.

Table 6 Computational efficiency under different hardware platforms

Platform	*Inference latency/ ms*	*End-to-end latency/ ms*	*Peak memory/ GB*	*Throughput/ (events·s^{-1})*
NVIDIA RTX 4070	41.3 ± 0.6	63.8 ± 1.4	0.84	24.2
RTX 3050 Laptop GPU	59.7 ± 1.2	86.4 ± 2.1	0.91	16.8
Jetson Orin NX 16 GB	77.9 ± 1.8	112.6 ± 2.9	1.06	12.8
Intel i7 CPU	126.4 ± 3.6	163.1 ± 4.2	1.32	7.9

As shown in Table 6, end-to-end feedback latency was 63.8 ms on the RTX 4070 and remained below 90 ms on the consumer RTX 3050 laptop GPU. On the Jetson Orin NX embedded platform, the complete processing chain required 112.6 ms with 1.06 GB peak runtime memory, demonstrating that classroom deployment does not require server-class hardware. Four concurrent performance streams increased peak memory to 2.1 GB and mean feedback latency to 69.5 ms, while eight streams required 3.4 GB and 88.2 ms, indicating approximately linear memory growth with acceptable latency under multi-stream operation.

3.2 Analysis of piano performance training outcomes

The effectiveness of training was evaluated by uniformly measuring rhythm deviation, note omission rate, dynamic coefficient of variation, and instructor evaluation before and after 8 weeks. The results are shown in Table 7.

Table 7 Comparison of performance quality before and after training (Mean ± SD)

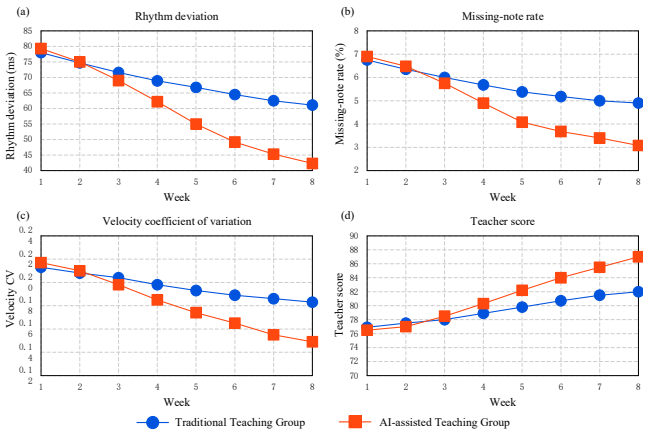
Group	*Phase*	*Rhythm deviation/ ms*	*Note omission rate/%*	*CV in dynamics*	*Teacher score/ 100*
Traditional	Before	78.4 ± 13.2	6.8 ± 1.6	0.214 ± 0.041	76.8 ± 5.4
Traditional	After	61.3 ± 11.4	4.9 ± 1.3	0.183 ± 0.036	82.1 ± 4.8
AI-assisted	Before	79.1 ± 12.7	6.9 ± 1.5	0.216 ± 0.043	76.5 ± 5.6
AI-assisted	After	42.7 ± 9.6	3.1 ± 1.0	0.149 ± 0.031	86.9 ± 4.2

As shown in Table 7, prior to training, the two groups showed comparable performance, whereas the AI-assisted group exhibited larger improvements after eight weeks. Rhythm deviation decreased from 79.1 to 42.7 ms, the note omission rate decreased from 6.9% to 3.1%, and the dynamics coefficient of variation decreased to 0.149. Mixed ANOVA revealed significant group × time interactions for rhythm deviation ($F_{1,46}$ = 34.8, p = 0.0001, η_p^2 = 0.431), note omission rate ($F_{1,46}$ = 22.6, p = 0.0002, η_p^2 = 0.329), dynamics variation ($F_{1,46}$ = 12.9, p = 0.0010, η_p^2 = 0.219), and instructor score ($F_{1,46}$ = 9.7, p = 0.0030, η_p^2 = 0.174). Holm-adjusted within-group comparisons in the AI-assisted group yielded changes of −36.4 ms in rhythm deviation (95% CI: −43.7 to −29.1; p = 0.0001, d = 2.10), −3.8 percentage points in omission rate (95% CI: −4.6 to −3.0; p = 0.0001, d = 1.95), −0.067 in dynamics variation (95% CI: −0.087 to −0.047; p = 0.0003, d = 1.43), and +10.4 points in teacher score (95% CI: 7.7–13.1; p = 0.0001, d = 1.65). Between-group change-score comparisons remained significant for all four outcomes (p = 0.0002–0.0026; d = 0.86–1.35).

Beyond immediate technical outcomes, the AI-assisted group showed higher behavioural engagement than the traditional group (90.7 ± 6.8% vs. 78.6 ± 8.9%, p < 0.001, d = 1.53) and lower cognitive workload (45.1 ± 7.6 vs. 54.2 ± 8.3 on NASA-TLX, p < 0.001). The system achieved an SUS score of 82.4 ± 6.7 and student satisfaction of 4.32 ± 0.41/5, while instructor acceptance reached 4.38 ± 0.36/5. At the four-week delayed assessment, rhythm deviation remained lower in the AI-assisted group (47.9 ± 10.1 ms vs. 63.0 ± 11.8 ms, p < 0.001), providing evidence of retention beyond the immediate post-test.

Figure 4 also shows the trends of these measurements during the 8-week training period. Figure 4 reveals that the downward trends of both groups were similar in the first 2 weeks. The rate for the AI-assisted instruction group rose from Week 3 onwards, while the rate of note omission continued to decline steadily after Week 5 and the teacher scores started to diverge from Week 6 onwards (86.9 points at Week 8, compared with 82.1 points at Week 8 for the traditional instruction group.

Figure 4 Changes in performance quality metrics during the training cycle (see online version for colours)

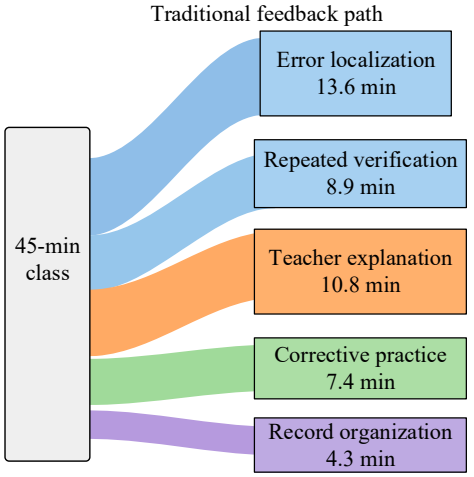


3.3 Results on classroom feedback efficiency

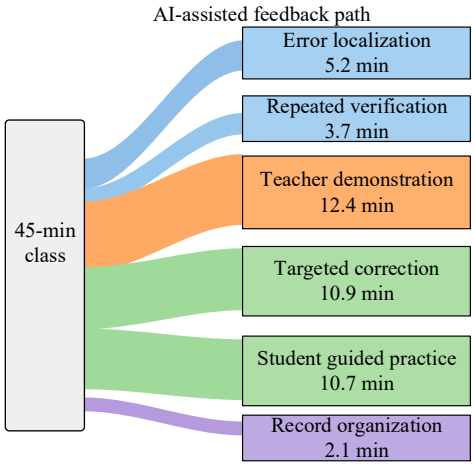
Classroom time allocation was obtained through frame-level coding of video recordings rather than instructor recall.

Sixteen matched 45-min lessons, eight from each teaching condition, were independently coded by two observers who were not involved in instruction and were unaware of the study hypothesis. Activity boundaries were defined using observable transitions: error localisation began when a performance segment ended and terminated when the target error was identified; repeated verification covered replay or repeated performance used to confirm the error; teacher explanation and demonstration began with explicit corrective instruction; targeted correction began when the student executed the prescribed correction; and record organisation covered logging and after-class task assignment. Disagreements exceeding 5 s were reviewed jointly. Inter-observer agreement reached ICC (2,1) = 0.94. Values in Figure 5 represent mean durations across the 16 coded sessions rather than measurements from a single lesson.

Figure 5 Comparison of time allocation for classroom feedback, (a) traditional (b) AI-assisted (see online version for colours)



(a)



(b)

The time flow diagram shows that the total class duration did not change, but the focus of feedback resources moved from low-level error correction to higher-value demonstration and correction sessions. The optimised priority weights achieved 88.6% agreement between the system-generated top-three feedback items and instructor selections, together with a 74.8% first-attempt correction rate. Under ±20% coefficient perturbations, teacher agreement remained between 86.9% and 88.7%, the correction rate ranged from 72.9% to 74.9%, and the Spearman correlation between perturbed and original feedback rankings remained between 0.93 and 0.98. Across rhythm-dominant, technique-dominant, and mixed-repertoire lessons, top-three agreement was 89.2%, 87.8%, and 88.4%, respectively. These results indicate that feedback ranking is not dependent on a narrowly tuned coefficient configuration.

A Sankey diagram comparing the two feedback pathways in a 45 minute class is shown in Figure 5. In the traditional process, the time used for error identification and repeated confirmation was the biggest (13.6 minutes, 8.9 minutes, respectively), and the time used for students' practice of correction was 7.4 minutes, and the time used for the explanation by the teacher was 10.8 minutes. Upon adding AI-assisted feedback, the system was able to automatically log events and aggregate errors, freeing up the teacher's time on the error identification task to 5.2 minutes and the time for repeated verification to 3.7 minutes. In turn, the time spent on demonstration and targeted correction rose to 12.4 minutes and 10.9 minutes respectively and the time spent on documentation was decreased from 4.3 minutes to 2.1 minutes. The time flow diagram (Figure 5) shows that the time allocated to the total class did not change but the focus of feedback resources moved from the use of low level error correction to high value demonstration and correction sessions.

3.4 Effectiveness of personalised training pathways

Student error profiles were developed using the occurrence, severity, and improvement rate of eight error categories. The number of K-means clusters was determined before interpretation by evaluating k = 2–6. The within-cluster sum-of-squares curve showed its principal elbow at k = 4, while the silhouette coefficient reached its maximum at 0.62 for k = 4, compared with 0.51, 0.56, 0.54, and 0.49 for k = 2, 3, 5, and 6, respectively. The Calinski–Harabasz index also reached its maximum value of 38.7 at k = 4. To assess initialisation sensitivity, K-means was repeated 50 times with different random seeds; the resulting partitions achieved a mean adjusted Rand index of 0.91 ± 0.04 relative to the consensus solution. Principal component analysis (PCA) was used only for visualisation in Figure 6, with the first two components explaining 71.4% of the total profile variance.

On this basis, four clusters were retained for subsequent training-path allocation:

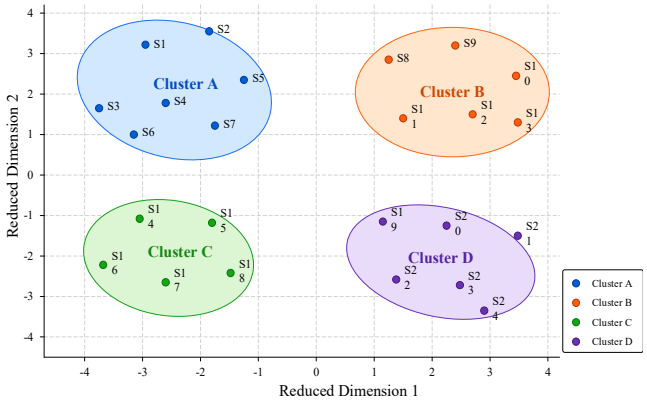
- Cluster A (7 students) includes primarily co-occurrence of rhythmic lag and missed notes
- Cluster B (6 students) includes primarily dynamic fluctuations and breaks from legato

- Cluster C (5 students) includes primarily deviations from the correct fingering paths and excessive wrist elevation
- Cluster D (6 students) includes primarily muddled pedal use and poor hand coordination.

Matching differentiated training paths (based on the clustering results) resulted in a 31.4%, 28.7%, 25.6%, and 22.9% reduction in the frequency of the dominant errors in the four clusters, respectively, for the next four weeks. In particular, the median value of the rhythm deviation in Cluster A was reduced from 84 to 51 milliseconds, while the percentage of deviation in the fingerings events in Cluster C was reduced from 12.8 to 8.3%.

As seen from Figure 6, there is minimal overlap between the boundaries of each of the clusters, suggesting that the error profiles are useful in the targeted allocations of tiered training tasks.

Figure 6 Clustering results of student error profiles (see online version for colours)



To further clarify the instructional contribution of each AI-assisted component, a counterbalanced component-removal evaluation was conducted within the 24 students in the AI-assisted group using matched technical passages. Each student completed three instructional conditions: automatic diagnosis only (D), diagnosis plus prioritised feedback (D + F), and the complete system with personalised training routes (D + F + P). The order of conditions was balanced across students, and repertoire difficulty was matched across conditions. First-attempt correction rate and subsequent error reduction were used as the primary evaluation metrics. The results are shown in Table 8.

Table 8 Component-removal evaluation of instructional modules

Condition	*First-attempt correction/%*	*Subsequent error reduction/%*
Diagnosis only (D)	58.3±8.1	14.2±5.6
Diagnosis + prioritised feedback (D+F)	68.7±7.4	21.8±6.1
Complete system (D + F + P)	75.4±6.9	29.6±6.5

As shown in Table 8, first-attempt correction improved progressively from 58.3 ± 8.1% under diagnosis only to 68.7 ± 7.4% after prioritised feedback was added, and further increased to 75.4 ± 6.9% when personalised training routes were included. Subsequent error reduction showed the same trend, increasing from 14.2 ± 5.6% to 21.8 ± 6.1% and then to 29.6 ± 6.5%. Repeated-measures analysis indicated significant incremental gains from D to D + F ($p = 0.004$) and from D + F to D + F + P ($p = 0.009$). These results suggest that automatic diagnosis provides the basic improvement, prioritised feedback contributes immediate corrective support, and personalised training routes yield additional consolidation of error reduction.

4 Discussion and conclusions

The results demonstrate that the proposed framework can connect event-level technical diagnosis with classroom feedback and differentiated practice in AI-assisted piano instruction. The main findings can be interpreted from four aspects.

1 Recognition performance gains arise from complementary multimodal evidence and score-conditioned reliability weighting. Audio features support pitch, timbre, and articulation-related errors; MIDI sequences contribute timing, duration, velocity, and pedal information; hand-pose features supply movement evidence for fingering-path and wrist-related errors. Removing the audio, MIDI, and pose branches reduced the F1-score by 5.4, 3.8, and 2.6 percentage points, respectively, while replacing the proposed reliability fusion with conventional attention caused a further 1.9-percentage-point reduction. The complete model achieved an F1-score of 92.2%, and five-fold subject-independent validation yielded 91.8±0.5%. External-institution and unseen-repertoire tests retained F1-scores of 90.5% and 89.8%. Compared with the stronger Transformer- and self-supervised baselines, the proposed model also retained the highest F1-score while requiring lower inference latency. These results indicate that performance is not dependent on repeated recordings from the same students or a single repertoire. The 41.3 ms inference latency also supports real-time classroom deployment.

2 Classroom outcomes indicate that technical improvement is accompanied by broader changes in learning. After eight weeks, rhythm deviation in the AI-assisted group decreased from 79.1 ms to 42.7 ms, note omission rate decreased from 6.9% to 3.1%, and instructor score increased from 76.5 to 86.9 points. Mixed ANOVA and Holm-adjusted comparisons confirmed significant group × time effects and meaningful effect sizes. Behavioural engagement was higher in the AI-assisted group, while cognitive workload was lower. Favourable SUS, satisfaction, and instructor-acceptance scores indicate that the system remained usable during repeated classroom application.

At the four-week delayed assessment, rhythm deviation remained lower than in the traditional group, showing that part of the technical gain persisted beyond the immediate post-test. Event-level localisation therefore directs practice toward specific measures, notes, and technical actions requiring correction.

3 Classroom feedback efficiency is reflected in the redistribution of instructional time and the contribution of individual system modules. Video coding of 16 matched lessons showed that AI assistance reduced the time required for error localisation and repeated verification, while increasing time for teacher demonstration, targeted correction, and guided practice. Inter-observer reliability confirmed the stability of the time-allocation measurements. The component-removal experiment further separated automatic diagnosis, prioritised feedback, and personalised training routes. First-attempt correction increased from 58.3% under diagnosis only to 68.7% after prioritised feedback was added and to 75.4% under the complete system, while subsequent error reduction increased from 14.2% to 21.8% and 29.6%. The system therefore supports repetitive diagnostic and organisational tasks and allows instructors to devote more classroom resources to explanation, demonstration, interpretation, and targeted correction.

4 Error-profile clustering provides an operational basis for differentiated practice allocation. The four-cluster solution was supported by the elbow criterion, a maximum silhouette coefficient of 0.62, the Calinski–Harabasz index, and stable partitions across 50 random initialisations. The resulting groups corresponded to dominant technical patterns involving rhythm and missed notes, dynamics and legato, fingering and wrist motion, and pedal use and hand coordination. Matching training routes to these profiles produced reductions of 31.4%, 28.7%, 25.6%, and 22.9% in dominant-error frequency. The high adjusted Rand index across repeated initialisations further indicates that these profiles were not artefacts of a single clustering seed. Student-level error distributions can therefore be translated into interpretable practice assignments rather than serving only as descriptive clusters.

Overall, the framework integrates multimodal acquisition, score-conditioned event reliability fusion, technical-error recognition, prioritised classroom feedback, and personalised practice within a unified computational pipeline. Its intended scope is technical performance diagnosis and formative classroom feedback. Automated outputs cover pitch, timing, fingering, dynamics, pedal use, and legato continuity. Phrasing, tone shaping, structural awareness, stylistic interpretation, and emotional expression remain under instructor judgment because these dimensions require higher-level musical context beyond the current event representation. Accordingly, recognition accuracy should be interpreted as technical-error recognition accuracy rather than a complete measure of piano artistry. Broader validation across geographic regions, age groups, instrument conditions, longer instructional periods, and higher-level expressive dimensions remains necessary.

Declarations

The data used to support the findings of this study are all in the manuscript.

The authors declare no competing interests.

References

Bontempi, P., Canazza, S., Carnovalini, F. and Rodà, A. (2023) 'Research in computational expressive music performance and popular music production: a potential field of application?', *Multimodal Technologies and Interaction*, Vol. 7, No. 2, p.15.

Burunat, I., Levitin, D.J. and Toiviainen, P. (2024) 'Breaking (musical) boundaries by investigating brain dynamics of event segmentation during real-life music-listening', *Proceedings of the National Academy of Sciences*, Vol. 121, No. 36, p.e2319459121.

Cho, W., Ha, T., Jeon, I., Jeon, J., Kim, T.K. and Woo, W. (2024) 'Temporally enhanced graph convolutional network for hand tracking from an egocentric camera', *Virtual Reality*, Vol. 28, No. 3, p.143.

Christodoulou, A.M., Lartillot, O. and Jensenius, A.R. (2024) 'Multimodal music datasets? Challenges and future goals in music processing', *International Journal of Multimedia Information Retrieval*, Vol. 13, No. 3, p.37.

Colafiglio, T., Ardito, C., Sorino, P., Lofù, D., Festa, F., Di Noia, T. and Di Sciascio, E. (2024) 'Neuralpmg: a neural polyphonic music generation system based on machine learning algorithms', *Cognitive Computation*, Vol. 16, No. 5, pp.2779–2802.

Deja, J.A., Štor, S., Pucihar, I., Weerasinghe, M., Balbin, R.M., Čopič Pucihar, K. and Kljun, M. (2025) 'ImproVisAR: designing augmented reality piano roll for teaching improvisation', *Virtual Reality*, Vol. 29, No. 3, p.140.

Dibenedetto, G., Sotiropoulos, S., Polignano, M., Cavallo, G. and Lops, P. (2025) 'Comparing human pose estimation through deep learning approaches: an overview', *Computer Vision and Image Understanding*, Vol. 252, p.104297.

Edwards, D., Dixon, S. and Benetos, E. (2023) 'Pijama: piano jazz with automatic midi annotations', *Transactions of the International Society for Music Information Retrieval*, Vol. 6, No. 1, pp.89–102.

Emerson, K. and Harrison, P.M. (2025) 'Multimodal datasets for studying expert performances of musical scores', *Transactions of the International Society for Music Information Retrieval*, Vol. 8, No. 1, pp.400–428.

Ghosh, A., Saini, A. and Barad, H. (2025) 'Artificial intelligence in governance: recent trends, risks, challenges, innovative frameworks and future directions', *AI and SOCIETY*, Vol. 40, No. 7, pp.5685–5707.

Kappan, M.M., Sandoval, E.B., Meijering, E. and Cruz, F. (2025) 'A survey on deep learning for 2D and 3D human pose estimation', *Artificial Intelligence Review*, Vol. 59, No. 1, p.32.

Kim, T., Kim, M.J. and Ahn, C.W. (2025) 'PF2N: periodicity–frequency fusion network for multi-instrument music transcription', *Mathematics*, Vol. 13, No. 11, p.1708.

Kyriakou, T., de la Campa Crespo, M.Á., Panayiotou, A., Chrysanthou, Y., Charalambous, P. and Aristidou, A. (2024) 'Virtual instrument performances (VIP): a comprehensive review', in *Computer Graphics Forum*, Vol. 43, No. 2, p.e15065.

Louro, P.L., Redinho, H., Malheiro, R., Paiva, R.P. and Panda, R. (2024) 'A comparison study of deep learning methodologies for music emotion recognition', *Sensors*, Vol. 24, No. 7, p.2201.

Moura, N., Vidal, M., Aguilera, A.M., Vilas-Boas, J.P., Serra, S. and Leman, M. (2023) 'Knee flexion of saxophone players anticipates tonal context of music', *NPJ Science of Learning*, Vol. 8, No. 1, p..

Mycka, J. and Mańdziuk, J. (2025) 'Artificial intelligence in music: recent trends and challenges', *Neural Computing and Applications*, Vol. 37, No. 2, pp.801–839.

Mycka, J. and Mańdziuk, J. (2026) 'Classification of piano performers with deep learning models', *Journal of Computational Science*, ID 102804.

Ostermann, F., Vatolkin, I. and Ebeling, M. (2023) 'AAM: a dataset of artificial audio multitracks for diverse music information retrieval tasks', *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2023, No. 1, p.13.

Özer, Y., Schwär, S., Arifi-Müller, V., Lawrence, J., Sen, E. and Müller, M. (2023) 'Piano concerto dataset (PCD): a multitrack dataset of piano concertos', *Transactions of the International Society for Music Information Retrieval*, Vol. 6, No. 1, pp.75–88.

Patil, S., Patil, R., Goudar, S., Sangani, S. and Goudar, R.H. (2025) 'Review on music emotion analysis using machine learning: technologies, methods, datasets, and challenges', *Discover Applied Sciences*, Vol. 7, No. 7, p.692.

Ramoneda, P., Jeong, D., Eremenko, V., Tamer, N.C., Miron, M. and Serra, X. (2024) 'Combining piano performance dimensions for score difficulty classification', *Expert Systems with Applications*, Vol. 238, p.121776.

Sujeesha, A.S. and Mala, J.B. (2024) 'Automatic music mood classification using multi-modal attention framework', *Engineering Applications of Artificial Intelligence*, Vol. 128, p.107355.

Volta, E. and Di Stefano, N. (2024) 'Using wearable sensors to study musical experience: a systematic review', *Sensors*, Vol. 24, No. 17, p.5783.

Woo, T., Park, W., Jeong, W. and Park, J. (2023) 'A survey of deep learning methods and datasets for hand pose estimation from hand-object interaction images', *Computers and Graphics*, Vol. 116, p.474–490.

Younesi Heravi, M., Jeong, I. and Jang, Y. (2026) 'A vision-based approach for human activity intensity estimation using kinematic features', *Journal of Computing in Civil Engineering*, Vol. 40, No. 6, p.4026092.