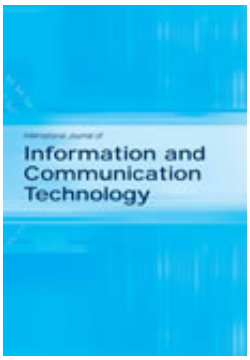


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Analysis of an intelligent piano music transcription model by deep reinforcement learning

Yan Hu, Jing Wang

DOI: 10.1504/IJICT.2026.10075844

Article History:

Received:	29 September 2025
Last revised:	07 November 2025
Accepted:	10 November 2025
Published online:	02 February 2026

Analysis of an intelligent piano music transcription model by deep reinforcement learning

Yan Hu and Jing Wang*

Humanities Quality Education Center,
University of Science and Technology,
Beijing 100083, China
Email: yah2@stir.ac.uk
Email: 15907437755@163.com
*Corresponding author

Abstract: To improve the accuracy of automatic piano music transcription in complex environments, a recognition system applicable to practical scenarios such as music education assistance and intelligent performance analysis was developed. First, audio features were extracted using Log-Mel spectrograms, combined with data augmentation and adaptive pitch normalisation to enhance model robustness. Second, a state-action modelling mechanism integrating a Transformer encoder with a multidimensional action space was constructed to precisely represent note content, rhythmic positions, and dynamics information. Finally, a primary policy and an auxiliary rhythm policy based on proximal policy optimisation (PPO) were designed, and a multidimensional reward function along with imitation learning signals were introduced to jointly optimise the note prediction strategy. Comparative experiments indicated that incorporating the multidimensional action structure and boundary auxiliary strategy significantly improved recognition accuracy. The proposed method achieves high-precision piano audio transcription with strong structural continuity.

Keywords: piano transcription; deep reinforcement learning; DRL; multidimensional action space; music sequence modelling; proximal policy optimisation; PPO.

Reference to this paper should be made as follows: Hu, Y. and Wang, J. (2026) 'Analysis of an intelligent piano music transcription model by deep reinforcement learning', *Int. J. Information and Communication Technology*, Vol. 27, No. 3, pp.18–35.

Biographical notes: Yan Hu is a Chief Scientist of the Omniverse Foundation, Hong Kong, General Manager of Hainan Dongqiang Cultural Group, and member of the China Association for Scientific Expedition, currently serves as a Lecturer (Intern) at the Humanities Quality Education Center, University of Science and Technology Beijing. He holds a DBA jointly awarded by UCASS and the University of Stirling, following earlier graduate and undergraduate training at the Central Conservatory of Music and Boston University. His research focuses on arts management, crisis management, and music technology.

Jing Wang is an executive council member of the Huqin Committee of the Hunan Ethnic Orchestra Society, and currently serves as a Lecturer at the Humanities Quality Education Center, University of Science and Technology Beijing. Trained at the China Conservatory of Music, she specialises in ethnic music education and musical performance. She has undertaken over ten

research projects and received nearly twenty professional honours, including the sole Gold Award at the 2011 28th Shanghai Spring Jiangnan Sizhu Competition and a finalist award at the 2015 Golden Bell Awards. Her representative achievements include two solo huqin recitals and original works such as MiaoZhai HuanGe.

1 Introduction

With the continuous development of artificial intelligence and audio signal processing technologies, music information retrieval (MIR) is progressively moving toward structured, real-time, and multimodal integration (Zhang, 2025). Among these, automatic music transcription (AMT), which connects audio signals with musical notation, has become a research focus in the field of music AI (Gao et al., 2023; Chen et al., 2022). The piano, as a polyphonic instrument with complex multi-track structures, contains rich rhythmic, dynamic, and layered information in its audio signals. Achieving high-precision piano music transcription not only supports applications in music education, performance assistance, and intelligent score generation but also provides a foundation for MIR, performance style analysis, and cross-modal generation (Guo, 2025).

Although deep learning methods have achieved significant progress in monophonic music transcription tasks in recent years, practical piano transcription still faces challenges such as insufficient accuracy in recognising multiple simultaneous notes, large errors in rhythm boundary prediction, and poor robustness under style transfer (Li, 2022; Jamshidi et al., 2024). On one hand, piano performances involve numerous chords, arpeggios, and sustained pedal effects, leading to substantial spectral overlap among notes and increasing recognition difficulty (Latif et al., 2023). On the other hand, background noise, reverberation, and individual variations in tempo and dynamics in real-world recordings further complicate modelling and limit the generalisation of traditional supervised models (Dai, 2023). Most existing approaches rely on fixed-label supervision and struggle to leverage dynamic feedback from the performance process, posing challenges for learning efficiency and policy adaptability.

Compared with traditional supervised learning methods, deep reinforcement learning (DRL) has stronger sequence modelling and feedback learning capabilities, especially suitable for task scenarios with complex state spaces and structured action spaces. In piano transcription, note recognition requires consideration of the acoustic features of the current audio frame, and relies on the rhythm and coherence of historical note sequences. DRL can optimise transcription strategies in dynamic environments through strategy learning mechanisms, combined with reward functions to automatically adjust recognition results, effectively improving performance in distinguishing complex note structures and rhythm boundaries. Meanwhile, DRL does not rely on strict label distribution and can continuously learn in weakly supervised or self-supervised data environments, improving generalisation ability and stability against noise.

To address these issues, this study proposes an intelligent piano music transcription method based on multi-strategy DRL. By constructing a joint state-action space modelling mechanism and employing a Transformer attention encoder to obtain context-aware representations, the model predicts note categories, onset-offset times, and dynamics using a multidimensional action structure. Additionally, a primary policy and a

boundary auxiliary policy based on proximal policy optimisation (PPO) were designed for collaborative optimisation. A reward function integrating pitch, rhythm, and structural coherence was employed to enhance the model's decision-making ability and structural restoration in complex scenarios. The structure of this study is as follows: Section 2 reviews related research; Section 3 introduces the model architecture and training methods in detail; Section 4 presents multiple experimental results and analyses; Section 5 discusses the proposed method and Section 6 concludes the study and outlines future research directions. The state action space joint modelling mechanism proposed here can significantly enhance the structural perception ability of multi note and rhythm boundaries. The multi-dimensional action output structure further enhances the linkage prediction ability of elements such as note start stop and intensity, thereby solving the challenge of identifying fuzzy boundaries between multiple notes and rhythm. The reinforcement learning strategy network effectively improves adaptability in different playing styles and complex backgrounds through a reward feedback-based strategy optimisation mechanism, and constructs a piano transcription model with more generalisation ability.

2 Literature review

Early AMT methods were primarily based on spectrogram feature extraction, fundamental frequency detection, and hidden Markov model (HMM) modelling (Zhai and Xu, 2022; Bhattarai and Lee, 2023). Although these approaches achieved certain success in simple melodies and monophonic music, their performance was limited for polyphonic instruments such as the piano. In recent years, with the development of deep learning, traditional models have gradually been replaced by data-driven architectures. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been widely used to model audio spectral and temporal sequence features (Lei and Liu, 2022). Large et al. (2023) investigated the Onsets and Frames v2 model, which independently models note onsets and sustain states, improving recognition accuracy under complex rhythms. Zhao et al. (2021) introduced residual convolutional structures combined with the connectionist temporal classification (CTC) loss function, enhancing model robustness against misalignment between notes and frame labels. Ananth et al. (2025) employed a dual-channel attention network to model pitch and the temporal axis in parallel, achieving strong performance in polyphonic transcription tasks. Meanwhile, transformer architectures have gradually been applied to audio modelling. Kamal et al. (2022) used transformers to capture long sequences, mitigating the failure of traditional RNNs in modelling long-term dependencies.

Building on these advances, some studies have explored incorporating DRL into music analysis tasks. Dadman and Bremdal (2024) applied policy optimisation to melody generation, introducing reward functions to control musical structure and style. Yaqoob et al. (2024) used an actor-critic strategy for audio event detection, improving the detection of event boundaries in complex backgrounds. Wang et al. (2024b) investigated note sequence generation based on PPO, achieving automatic learning of rhythmic structures on synthetic data.

Currently, two main challenges remain: reinforcement learning methods exhibit poor adaptability to polyphonic structures and sustained pedal effects, and the design of state representations and action spaces is complex; additionally, reward functions struggle to

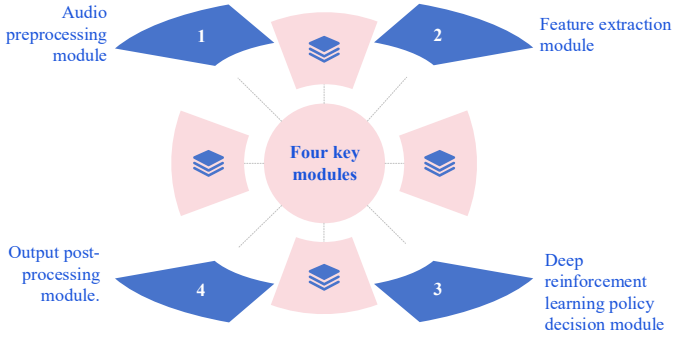
quantify rhythmic continuity and expressive performance, leading to unstable policies and difficulty in training convergence. Most existing models rely on single-policy optimisation and lack collaborative modelling of rhythmic boundaries and overall musical structure. To address these limitations, this study proposes an integrated policy learning framework that combines a transformer encoder, composite action modelling, and multidimensional reward structures, aiming to improve recognition accuracy and model generalisation for piano audio under complex performances and real-world conditions.

3 Multi-policy fusion DRL piano transcription model design

3.1 System framework

The overall structure of the proposed intelligent piano music transcription model is illustrated in Figure 1. The system consists of four key modules: an audio pre-processing module, a feature extraction module, a DRL policy decision module, and an output post-processing module. The system takes raw piano audio as input and outputs structured musical instrument digital interface (MIDI) transcriptions, including note categories, onset-offset times, and dynamics information.

Figure 1 Overall architecture of the intelligent piano music transcription model (see online version for colours)



Based on Figure 1, the audio signal is transformed using short-time Fourier transform (STFT) and a Log-Mel filter bank to extract its time-frequency feature map. The feature map is then fed into a deep encoder, where multiple convolutional layers and a transformer encoder extract contextual correlation features to form the current state representation. On this basis, the state vector is passed to the DRL policy network for action decisions, outputting composite actions such as note predictions and multi-note onset-offset boundaries. The policy network is built on an actor-critic architecture, which continuously optimises note recognition strategies through interactions with the environment. This module adopts a typical actor-critic architecture and includes the following three main sub-networks. Policy network: based on bidirectional GRU and multi-layer perceptron structure, it inputs the current state vector s_t and outputs a multidimensional action vector $a_t = (n_p, \delta_{on}, \delta_{off}, v, \theta)$. It is used to predict the note category, start and end time offset, intensity, and boundary control of the current frame.

value network (V_ϕ): the structure and policy network share some encoder parameters, which is used to evaluate the expected long-term reward $V_\phi(s_t)$ of the current state and assist in optimising the policy direction. Auxiliary policy (π_{aux}): used to capture fine-grained rhythm boundary signals, the network structure introduces attention mechanism to focus on the start and end time frames of notes, and outputs rhythm boundary suggestion values and their confidence levels. The training of each sub-network is achieved collaboratively through shared state representation and independent output of decision results. The final action selection combines the main strategy and auxiliary strategy through a strategy fusion mechanism (soft fusion or gating replacement) to improve the accuracy of boundary determination and the robustness of the strategy.

Environment feedback is evaluated using a carefully designed reward function, incorporating pitch accuracy, rhythmic precision, and performance continuity. Finally, the output actions are mapped into structured MIDI events through a post-processing module, where redundancy elimination and temporal correction are applied to generate more natural and fluent transcription results.

3.2 Audio feature representation and pre-processing optimisation

A spectrogram extraction method based on a Log-Mel filter bank was employed. This approach is perceptually closer to the human auditory frequency response and is well suited for capturing low-frequency harmonics, subtle timbral variations, and nonlinear dynamics in piano audio (Shang, 2022; Ji et al, 2023). The raw audio signal $x(t)$ is transformed into a complex spectrogram using STFT, as shown in equation (1):

$$X(n, \omega) = \sum_{m=0}^{M-1} x(m) \cdot w(n-m) \cdot e^{-j\omega m} \tag{1}$$

where $w(n)$ denotes the Hamming window function and M is the window length. The spectrogram is then mapped to the Mel filter bank and logarithmically compressed, yielding equation (2):

$$S_{\text{log-mel}}(n, m) = \log\left(\sum_{k=1}^{K} |X(n, k)|^2 \cdot H_m(k)\right) \tag{2}$$

where $H_m(k)$ represents the response of the m^{th} Mel filter and K is the number of frequency channels. The resulting two-dimensional feature map $S_{\text{log-mel}} \in \mathbb{R}^{T \times F}$ serves as the primary state representation for subsequent modelling.

To enhance model robustness under diverse performance styles and complex acoustic backgrounds, a series of data augmentation strategies were introduced, including time-frequency masking (SpecAugment) and time-stretching (Liu and Zhu, 2025; Dai, 2022). SpecAugment applies zero masks to random time and frequency segments of the feature map, simulating local note loss or occlusion (Colafiglio et al., 2024). Time-stretching adjusts playback speed to simulate performance style variations, with a perturbation factor $\alpha \in [0.9, 1.1]$, producing training samples of varied tempos as $x'(t) = x(\alpha t)$ (You, 2024).

This study also proposed an adaptive pitch normalisation method to mitigate pitch drift caused by different performers or recording devices. The method is based on global spectral centroid and dynamic range normalisation, linearly shifting the audio spectrum in the frequency domain so that its main frequency peak aligns with the dataset's reference

mean μ_f. If the principal frequency of a sample is f_p, the pitch offset is defined in equation (3):

$$\Delta f = \mu_f - f_p \tag{3}$$

The spectrogram is then shifted by Δf frequency bins, resampled, and reconstructed, effectively improving robustness against pitch variations without introducing additional annotation costs, thereby enhancing generalisation.

3.3 Innovations in state space and action space modelling

To enable the DRL policy network to capture the multidimensional structural features of audio signals, a state space with strong semantic representation and explicit contextual dependency is required. In addition, a multidimensional action space must be designed to express composite decisions, supporting precise note recognition and boundary prediction. This overcomes the limitations of traditional music classification tasks in terms of action granularity and continuity modelling.

The state space is designed by integrating the two-dimensional Log-Mel spectrogram with contextual vectors (Liang and Pan, 2023). At each time step t, the input state s_t consists of the spectrogram matrix $S_{\text{log-mel}}^{(t)} \in \mathbb{R}^{W \times F}$ within the current time window, along with preceding contextual information. Contextual modelling is achieved through a transformer encoder with multi-head attention, capturing long-term temporal dependencies within the spectrogram sequence. The state vector is defined in equation (4):

$$s_t = \text{Transformer}\left(\left\{S_{\text{log-mel}}^{(t-k)}, ..., S_{\text{log-mel}}^{(t)}\right\}\right) \tag{4}$$

where k denotes the context window length.

The action space design breaks away from the conventional 'single-classification' or 'single-note output' paradigm in reinforcement learning by adopting a multidimensional action vector to represent composite decision-making intent. This more naturally simulates human auditory perception of holistic note structures (Wang, 2025; Liang, 2023). At each time step, the action a_t is defined as a 5-tuple, shown in equation (5):

$$a_t = \left(n_p, \delta_{\text{on}}, \delta_{\text{off}}, v, \theta\right) \tag{5}$$

where $n_p \in \{0, 1, \ldots, N\}$ denotes the predicted note index (including silence), $\delta_{\text{on}} \in \mathbb{R}^+$ and $\delta_{\text{off}} \in \mathbb{R}^+$ represent the onset and offset time offsets relative to the current frame, $v \in [0, 1]$ is the normalised note velocity, and $\theta \in \{0, 1\}$ is a note boundary flag used to assist policy convergence during training.

Although multidimensional action vectors can more naturally express complex musical events, their higher dimensions do bring about an increase in computational complexity. To control the training difficulty of the strategy network, this study conducts compression modelling in the selection of action dimensions, retaining only the five elements that have the most significant impact on the final transcription quality, and controlling the overall action dimension within 10. The strategy network adopts parameter sharing mechanism and weight freezing strategy to unify feature encoding for

different sub action channels, effectively alleviating training instability and overfitting problems.

3.4 Design of DRL-based transcription strategy

The transcription task of piano music has an inherent sequential decision-making property. Each action at a given frame depends not only on the current state but also on historical outputs (Wang et al., 2024a). To model such dynamic decision-making problems, a policy optimisation mechanism based on DRL is introduced. The actor-critic framework serves as the core, where the policy network and the value function network are decoupled. The policy network generates an action vector at each state, defined as $\pi_\theta(a_t|\ s_t)$. The value function network estimates the expected return of the state, $V_\phi(s_t)$, which guides directional policy improvement and strengthens the learning of long-term dependencies.

To improve the interpretability of learning signals and ensure stable training, a composite reward function is designed by integrating perceptual quality and performance structure (Ferreira et al., 2023). This function evaluates not only pitch recognition accuracy but also rhythmic alignment and continuity of dynamics. The overall reward function is expressed as equation (6):

$$R_t = \lambda_1 \cdot r_{\text{pitch}}(a_t) + \lambda_2 \cdot r_{\text{rhythm}}(a_t) + \lambda_3 \cdot r_{\text{contiguity}}(a_t) \tag{6}$$

r_{pitch} measures pitch-level alignment between transcription results and ground-truth notes using Levenshtein distance. r_{rhythm} evaluates the inverse mean-square error of onset and offset deviations. $r_{\text{contiguity}}$ defines a continuity score function based on temporal and dynamic consistency across notes. The coefficients λ_i control the weights of different reward dimensions, and cross-validation determines their optimal configuration.

The design of the reward function fully considers the subjective factors of music perception. For example, the tolerance setting for rhythm boundary deviation in the r_{rhythm} reward is based on the subjective perceivable threshold of human rhythm error (about 20 ms) in Bonnet et al. (2024), while the $r_{\text{contiguity}}$ coherence score refers to the importance of performance smoothness and rhythm fluency on listening sensation in music psychology (Di Stefano et al., 2024). By integrating these perceptual factors, the reward signal can be closer to the actual auditory experience, enhancing the rationality and generalisation ability of strategy learning.

To accelerate early policy convergence, imitation learning signals from expert performance data (MIDI labels) are introduced. A behaviour cloning loss is added as an auxiliary training objective (Phatnani and Patil, 2024). Given the expert action a_t^*, the loss is defined as equation (7):

$$\mathcal{L}_{\text{BC}} = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\left\|\pi_0(s_t) - a_t^*\right\|^2\right] \tag{7}$$

For policy optimisation, PPO is applied as the main strategy. PPO balances stability and convergence efficiency. The objective function is defined as equation (8):

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right] \tag{8}$$

where $r_t(\theta) = \dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the generalised advantage estimator. An auxiliary policy network is introduced for rhythmic boundary refinement. With an attention mechanism, it focuses on the temporal distribution of note onsets and offsets. A rhythm-boundary discriminator $D_\psi(s_t)$ is constructed to strengthen perception and correction in ambiguous boundary regions. The main and auxiliary policies are integrated through multi-policy fusion. In soft fusion, outputs are combined with weighted averaging as shown in equation (9):

$$a_t^{\text{final}} = \alpha \cdot a_t^{\text{main}} + (1-\alpha) \cdot a_t^{\text{aux}} \tag{9}$$

In hard substitution, a boundary confidence gating mechanism switches policies. Based on the auxiliary network's confidence score γ_t, the auxiliary output overrides the main policy when necessary.

To enhance the adaptability of strategy fusion, this study designs a heuristic strategy selection mechanism. Firstly, the boundary confidence level γ_t output by the auxiliary strategy determines the strategy switching gate. If γ_t is higher than the set threshold, the hard substitution mode is enabled to improve the boundary accuracy. On the contrary, soft fusion is used to maintain the continuity of the strategy. In the future, a learnable mechanism of strategy fusion weights can be further introduced, using meta learning or strategy attention networks to dynamically adjust the α value to adapt to more diverse music contexts.

3.5 Model training and optimisation procedure

The overall training process consists of two stages. In the first stage, imitation learning pre-trains the policy network by using existing MIDI data as expert demonstrations. This stage establishes an initial policy. In the second stage, reinforcement learning fine-tunes the policy through interaction with the environment, which further improves generalisation on real piano audio.

During pre-training, behaviour cloning minimises the mean squared error between policy outputs and expert actions. Given the training sample state s_t, expert action a_t^*, and predicted output $\pi_\theta(s_t)$, the loss is defined as equation (10):

$$\mathcal{L}_{\text{pre}} = \frac{1}{T}\sum_{t=1}^{T} \left\| \pi_\theta(s_t) - a_t^* \right\|^2 \tag{10}$$

To ensure consistency between the two-stage training, the reinforcement learning stage continues to use the same state representations and MIDI labels as the imitation learning stage, and the empirical sampling environment maintains consistent data sources. In order to alleviate the problem of 'policy collapse' or 'catastrophic forgetting' in the process of policy transfer, a mixed loss function is introduced in the early stage of policy fine-tuning, and the PPO objective is jointly trained with imitation learning loss to smooth the policy transfer process. Its positive effects on performance preservation and policy stability are verified in experiments.

Once the policy acquires basic note recognition ability, reinforcement learning begins. The system collects experience through environment interaction and optimises the policy with reward signals. To improve sampling efficiency and stability, an

experience replay mechanism stores state-action-reward sequences in a memory buffer M, and mini-batches are sampled for training.

Several control techniques are applied to stabilise and improve convergence. An entropy regularisation term is added to the policy objective to encourage sufficient exploration and avoid premature convergence to local optima. The modified PPO objective is expressed as equation (11):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PPO}} - \beta \cdot \mathbb{E}_t \left[\mathcal{H}\left(\pi_\theta\left(\cdot|s_t\right)\right) \right] \tag{11}$$

where $\mathcal{H}(\cdot)$ represents the entropy of the policy distribution, and β adjusts exploration strength. During parameter updates, gradient clipping constrains the maximum gradient norm to δ, which prevents instability from abnormal gradient fluctuations. In addition, an adaptive learning rate scheduling strategy adjusts the learning rate based on the dynamic variation of policy loss. This strategy effectively balances convergence speed and performance improvement.

4 Performance validation and comparative experiments of the multi-policy DRL piano transcription model

4.1 Experimental setup and evaluation metrics

The experiments are conducted on the Multitrack Alignment and Synchronous TRanscription of piano (MAESTRO) dataset. Provided by the Google Magenta team, this dataset contains over 200 hours of high-quality piano performances with aligned audio and MIDI labels. It covers a wide range of musical styles and techniques, including classical, romantic, and modern works, and supports fine-grained note-level transcription tasks. The experimental samples are selected from different composers' works in the dataset. They are further divided into subsets based on performance style, tempo range, and background complexity to simulate diverse real-world transcription scenarios. All audio samples are formatted as 44.1 kHz, 16-bit, mono signals. The training-to-test split is set at 8:2. The test set ensures heterogeneity in both styles and performers compared with the training set, which allows validation of the model's generalisation ability.

This study selects approximately 3500 pieces of music (2–15 s in duration) from MAESTRO and constructs three test subsets according to the following rules:

1 Style dimensions include classical (approximately 35%), Baroque (approximately 25%), jazz (approximately 20%), and modern pop (approximately 20%).

2 The speed dimension includes < 80 beats per minute (BPM) (slow), 80–120 BPM (medium), and > 120 BPM (fast).

3 The background complexity is simulated by adding environmental noise and reverberation.

Each subset ensures a balance between style and speed, ensuring diversity and representativeness in testing.

After preliminary tuning, the training parameters are set as follows: the initial learning rate is 0.0003, the batch size is 64, and each reinforcement learning iteration includes 2,048 environment interaction steps. The replay buffer size is 10,510^5,105. The

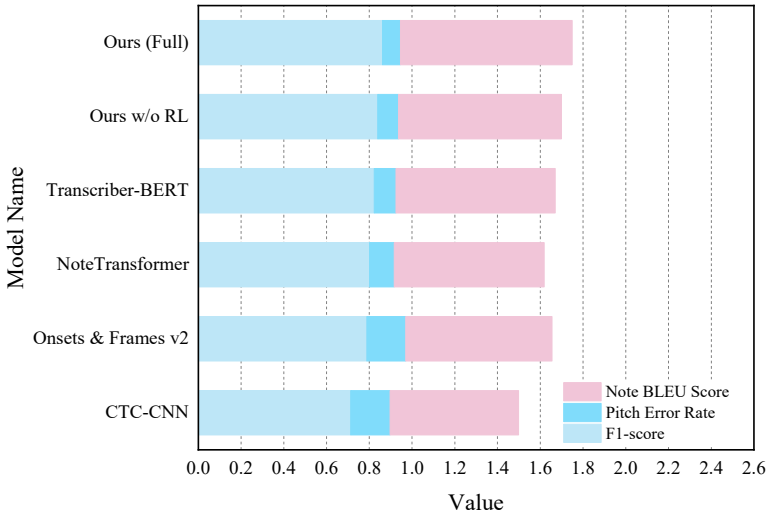
model training follows a five-stage joint strategy: the first two stages use behaviour cloning for pre-training, and the remaining stages apply reinforcement learning. The policy network is updated using PPO, while the value network is synchronised every two iterations.

4.2 Comparative evaluation with multiple models

To systematically evaluate the proposed multi-policy DRL piano transcription model across key performance dimensions, a comparative study with multiple baseline methods is conducted. Four representative approaches are selected as benchmarks. Two versions of the proposed model – one without reinforcement learning (ours w/o RL) and the full model (ours) – are also included to complete the comparison framework. The evaluated methods are: CTC-CNN, Onsets & Frames v2, NoteTransformer, Transcriber-BERT, ours w/o RL, and ours (full).

All models are trained and tested on the MAESTRO dataset, using the same train/test split and evaluation metrics to ensure fairness. Each model runs on the same hardware platform with its optimal public configuration and hyperparameter settings. The evaluation covers several dimensions: note recognition F1-score, pitch error rate (PER), onset/offset deviation error (ms), note bilingual evaluation understudy (BLEU) score, and average inference latency. Together, these metrics assess accuracy, structural fidelity, and real-time performance. The results are presented in Figures 2 and 3.

Figure 2 Comparative results of different models in terms of F1-score, PER, and note BLEU score (see online version for colours)

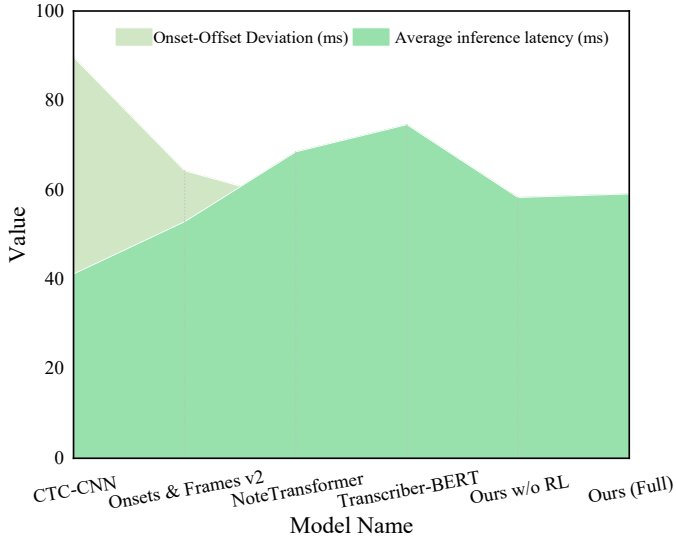


From Figures 2 and 3, it can be observed that the complete model proposed in this study outperforms all baseline methods across all key performance indicators. The F1-score reaches 0.862, representing a substantial improvement over the traditional CTC-CNN model. Meanwhile, the PER decreases to 8.4%, demonstrating a significant advantage in note recognition accuracy. The improvement in the Note BLEU score (reaching 0.803)

further indicates that the proposed method better preserves structural consistency and rhythmic continuity within musical segments.

In terms of rhythmic alignment, the proposed model achieves a note onset/offset deviation error of 38.5 ms, which is clearly superior to CNN-based static modelling methods and onsets and frames. This result verifies the effectiveness of the multidimensional action space and auxiliary rhythm strategy in structural reconstruction. Although the inference latency is slightly higher than that of the CTC-CNN model, the proposed approach maintains strong real-time performance while achieving higher accuracy, demonstrating its potential for practical deployment.

Figure 3 Note onset/offset deviation error (ms) and average inference latency (ms) (see online version for colours)



4.3 Ablation studies and validation of module effectiveness

To further verify the contribution of each key module to the overall performance of the proposed piano transcription model, multiple ablation experiments are designed. In each case, one core component of the model is removed, and performance is evaluated under identical training configurations and dataset partitions. Four model variants are included:

1 removing the auxiliary strategy network and retaining only the main strategy for note prediction

2 removing the rhythm continuity term from the joint reward function to examine its impact on output quality

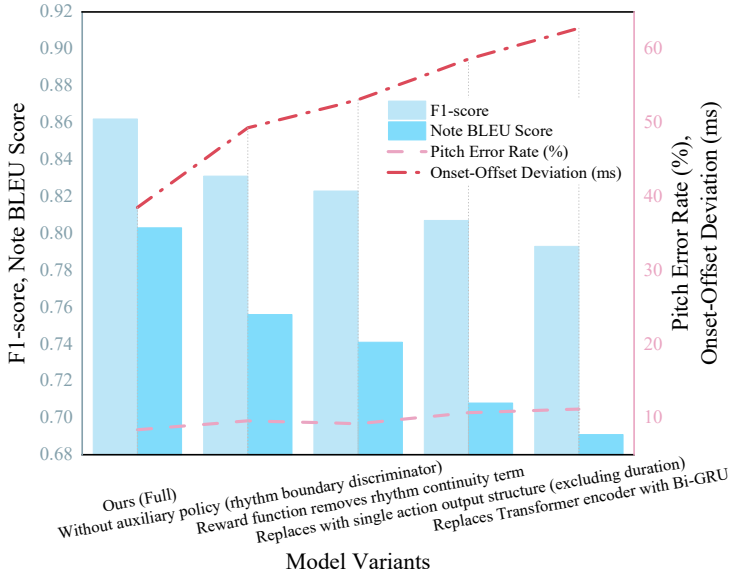
3 replacing the multidimensional action space with a single-note classification structure, ignoring onset/offset and velocity prediction

4 replacing the transformer encoder with a bidirectional GRU to assess the effect of temporal dependency modelling on strategy accuracy.

Each experiment is repeated three times, and average results are reported. The results are shown in Figure 4.

Based on Figure 4, removing the auxiliary strategy network causes a marked drop in both F1-score and BLEU structural scores. The onset/offset deviation increases by nearly 11 ms, highlighting the crucial role of the rhythm boundary discriminator in time-structure modelling. This component effectively alleviates boundary ambiguity and failures in handling legato passages. When the rhythm continuity term is excluded from the reward function, the note BLEU score drops to 0.741, and rhythm deviation increases by approximately 15 ms. This indicates that the absence of rhythmic structural supervision leads the strategy to favour short-term optimisation while neglecting overall rhythmic flow. Using a single-action output structure significantly weakens the model's ability to capture note duration and velocity, resulting in performance degradation in both pitch prediction and rhythm control. This finding further validates the structural advantages of the multidimensional action space in complex music sequence modelling. Replacing the Transformer with a bidirectional GRU preserves basic sequential modelling ability, but fails to capture long-term dependencies and global attention relationships. Consequently, the F1-score falls to 0.793, and the BLEU score for note sequences declines substantially.

Figure 4 Results of the module ablation experiments (see online version for colours)



4.4 Impact of reward function design on performance

To assess the actual contribution of each reward component to the policy optimisation process, four controlled experiments are designed, each enabling different reward dimensions. The configurations are as follows:

1 using only the pitch-matching term

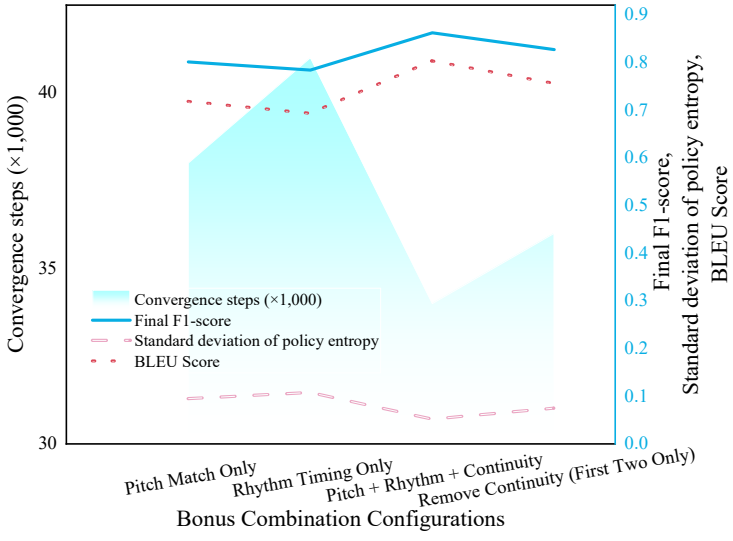
2 using only the onset/offset timing term

3 adding the note continuity term

4 using the complete joint reward structure.

All experiments are conducted under identical initial weights and training data conditions. For each case, convergence speed within 50 k steps, final F1-score, and fluctuation during training (measured by the standard deviation of policy entropy) are recorded. The results are shown in Figure 5.

Figure 5 Statistical analysis of the impact of reward function combinations on policy performance (see online version for colours)



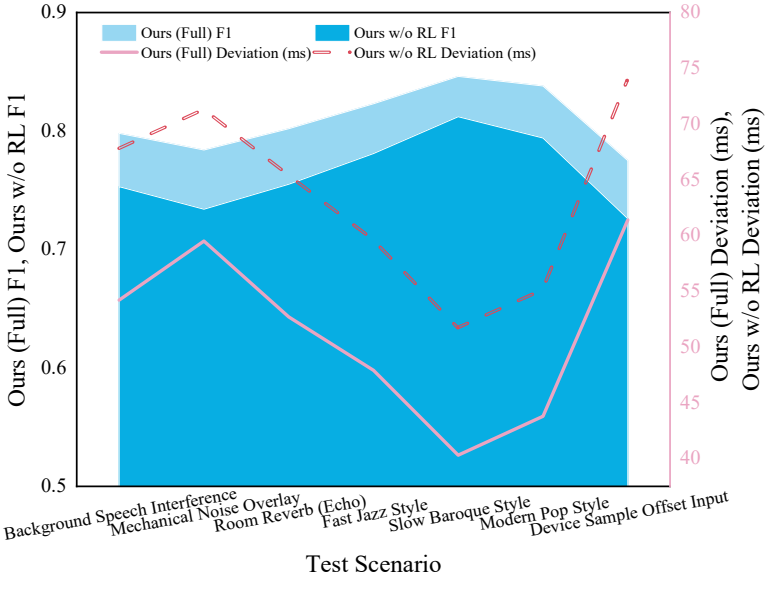
As shown in Figure 5, relying solely on pitch or rhythm rewards guides the policy toward learning basic recognition capabilities, but results in weaker overall structural reconstruction ability and stability. When the note continuity reward is introduced, the policy converges faster (with the number of steps reduced by about 10%), and the entropy standard deviation decreases significantly, indicating reduced volatility. Meanwhile, the improvement in the BLEU score confirms its effectiveness in preserving musical coherence.

4.5 Robustness evaluation

To evaluate the model's adaptability in real-world environments, several non-ideal scenario simulations are designed. These included background noise interference, style diversity variations, and device sampling mismatches. All tests used samples from the MAESTRO dataset, with data augmentation and parameter adjustments applied for simulation. For the noise environment test, background speech, mechanical noise, and room reverberation were added to emulate open-space performance conditions. For the style diversity test, subsets of different genres – jazz, Baroque, and pop – were selected to assess adaptability to tempo and expressivity variations. For the device sampling mismatch test, audio inputs were generated at different sampling rates (32 kHz, 22 kHz), bit depths, and frequency ranges to mimic discrepancies across recording devices.

Performance was evaluated using two metrics: accuracy (F1-score) and rhythm deviation (onset-offset deviation). Both the complete model (ours full) and the supervised ablation version (ours w/o RL) were compared. Results are presented in Figure 6.

Figure 6 Robustness evaluation results under non-ideal test conditions (see online version for colours)



As illustrated in Figure 6, the proposed model consistently outperforms the ablation version without reinforcement learning across all complex test conditions, with particularly stable performance in rhythm deviation control. In noisy environments, the average F1-score improves by about 4.5%, and rhythm errors decrease by more than 10 ms, demonstrating that the policy network can effectively adapt to input fluctuations through environmental feedback. In the style adaptability tests, when facing variations in tempo and expressivity, the reinforcement learning strategy maintained stronger rhythm stability, with BLEU structural consistency also improving accordingly.

This study designs two sets of refined ablation experiments to remove force prediction and boundary marker information, respectively, to verify the contributions of each dimension in the action space. The experimental results are shown in Table 1.

Table 1 Experimental results of refined ablation of action space

Model variant (action space dimension)	*F1-score*	*Note BLEU*	*Starting and ending time offset (ms)*
Ours (Full)	0.862	0.803	38.5
Remove the dimension of force	0.843	0.751	42.7
Remove boundary markers (without auxiliary strategy)	0.835	0.768	50.1

Table 1 show that the BLEU score decreased from 0.803 to 0.751 after the absence of the strength dimension, indicating its crucial role in maintaining the consistency of musical

sentence structure. After removing the boundary markers and corresponding auxiliary strategies, the start stop time offset error significantly increased from 38.5 ms to 50.1 ms, verifying the optimisation effect of rhythm assisted mechanism on temporal modelling.

To further evaluate the adaptability of soft fusion and hard substitution strategies in different music styles, this study applies two fusion methods to fast-paced jazz and rhythmic-balanced Baroque music segments for testing. The results are shown in Table 2.

Table 2 Transcriptional performance of different fusion strategies under different music styles

Styles	*Strategic integration approach*	*F1 score*	*Boundary accuracy*	*PER*
Jazz	Soft fusion	86.20%	81.50%	11.80%
Jazz	Hard substitution	85.90%	86.20%	11.30%
Baroque	Soft fusion	90.70%	84.40%	9.60%
Baroque	Hard substitution	87.30%	81.90%	10.10%

The results in Table 2 indicate that in fast-paced jazz passages, the hard substitution strategy improves the average rhythm boundary accuracy by about 4.7%, while in the Baroque style with a steady rhythm; the soft fusion strategy can increase the overall transcription F1 score by 3.1%.

5 Discussion

From the perspective of modelling note and rhythm structures, represented by Wei et al. (2022), they improved transcription accuracy through harmonic expansion convolution and frequency grouping RNN, but still relied on supervised labels and static modelling mechanisms. In contrast, the strategy network and action space design in this study enable the model to actively decide on the start and end, intensity, and rhythm boundaries of musical notes, thereby enhancing its adaptability to complex polyphonic instruments. From the perspective of generative and weakly supervised approaches, Marták et al. (2022) proposed treating transcription as a conditional generative task and achieved significant improvements, but it mainly focused on generative ability and lacked real-time decision feedback mechanisms. In contrast, this study integrates the 'decision-feedback' mechanism into the transcription process through reinforcement learning, thereby improving the model's ability to adjust strategies and restore structure accurately. From the perspective of the application of reinforcement learning in music tasks, Peter (2023) used reinforcement learning to achieve symbol music alignment tasks, verifying the potential of reinforcement learning in note structure recognition. Although its application scenarios are slightly different, it provides methodological support for the strategy optimisation framework in this study. Based on the above comparison, the highlight of this study is the establishment of a state action decision-making system directly facing transcription tasks, which fills the gap in the strategy feedback stage of traditional supervised learning. This study refines the action space design to multi-dimensional outputs such as note categories, start and end times, intensity, and boundary markers, making structured outputs more diverse. The study introduces strategy fusion mechanism and multidimensional reward function to enhance the performance of the model in rhythm boundary discrimination and performance style transfer. These mechanisms have been validated in experiments to enhance performance. However, this

study also has certain limitations. Firstly, although the experiment covers various styles and background environments, it has not yet been validated in a large number of multi instrument, extreme improvisation, or live recording scenarios. Secondly, although reinforcement learning mechanisms introduce structural feedback, policy training still heavily relies on labelled data and manually designed reward functions, and there is still room for improvement in terms of automation and weak supervision. Finally, although the real-time performance and resource consumption of the model are within an acceptable range, there is still room for optimisation in larger scale audio input or edge device deployment scenarios. Future research can be conducted in the following directions:

1 Introduce multimodal data (such as performance videos and finger detection) to enhance state representation and strategic decision-making capabilities.

2 Explore self-supervised or semi supervised reinforcement learning frameworks to reduce reliance on high-quality labels.

3 Extend the model to multi instrument ensemble and live multi-source recording environments to verify the generalisation ability of structured strategies in more complex music ecosystems.

In summary, this study enriches the methodological perspective of piano automatic transcription, and provides theoretical and practical references for the development of intelligent music understanding systems towards decision-making and structural perception.

6 Conclusions

This study proposes an intelligent piano music transcription model based on multi-policy DRL, establishing a state-action modelling framework that integrates Log-Mel spectrograms, a transformer encoder, and a multidimensional action space. For policy learning, a joint reward function was designed and imitation learning signals were introduced to guide efficient convergence of the PPO strategy under the actor-critic architecture. An auxiliary rhythm boundary discriminator and policy fusion mechanism further enhanced rhythm recognition and boundary alignment. Empirical studies conducted on the MAESTRO dataset demonstrate that the proposed model outperforms existing representative approaches in multiple dimensions, including F1-score, PER, rhythm deviation, and structural BLEU score. In robustness tests under complex conditions, the reinforcement learning strategy exhibited stronger noise resistance and style adaptability, confirming its potential for practical deployment. Despite the promising results on a single-instrument dataset, limitations remain. The model has not yet been validated in multi-instrument polyphonic scenarios, and its handling of expressive parameters, (e.g., pedal use and dynamic variations) is still incomplete. Future work will focus on expanding multimodal input structures. This includes incorporating fingering videos and score information. In addition, reward function designs that capture deeper awareness of musical structure will be explored. These improvements aim to enhance the model's performance in more complex music understanding tasks.

Data availability statement

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

Declarations

The authors declare no conflicts of interests.

References

Ananth, P., Kothandaraman, M. and Ishwarya, V.S. (2025) 'Multi-channel audio enhancement using dual-stream encoders with attention mechanisms and spatial discrimination GAN', *Circuits, Systems, and Signal Processing*, Vol. 44, pp.5945–5989.

Bhattarai, B. and Lee, J. (2023) 'A comprehensive review on music transcription', *Applied Sciences*, Vol. 13, No. 21, p.11882.

Bonnet, P., Bonnefond, M. and Kösem, A. (2024) 'What is a rhythm for the brain? The impact of contextual temporal variability on auditory perception', *Journal of Cognition*, Vol. 7, No. 1, p.15.

Chen, S., Zhong, Y. and Du, R. (2022) 'Automatic composition of Guzheng (Chinese Zither) music using long short-term memory network (LSTM) and reinforcement learning (RL)', *Scientific Reports*, Vol. 12, No. 1, p.15829.

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.

Dadman, S. and Bremdal, B.A. (2024) 'Crafting creative melodies: a user-centric approach for symbolic music generation', *Electronics*, Vol. 13, No. 6, p.1116.

Dai, L. (2022) 'Analysis of two-piano teaching assistant training based on neural network model sound sequence recognition', *Computational Intelligence and Neuroscience*, Vol. 2022, No. 1, p.5768291.

Dai, S. (2023) *Towards Artificial Musicians: Modeling Style for Music Composition, Performance, and Synthesis via Machine Learning*, Diss. Stanford University, Vol. 1, pp.1–32.

Di Stefano, N., Lo Presti, D., Raiano, L., Massaroni, C., Romano, C., Schena, E., Leman, M. and Formica, D. (2024) 'Expressivity attributed to music affects the smoothness of bowing movements in violinists', *Scientific Reports*, Vol. 14, No. 1, p.22267.

Ferreira, P., Limongi, R. and Fávero, L.P. (2023) 'Generating music with data: application of deep learning models for symbolic music composition', *Applied Sciences*, Vol. 13, No. 7, p.4543.

Gao, W., Zhang, S., Zhang, N., Xiong, X., Shi, Z. and Sun, K. (2023) 'Generating fingerings for piano music with model-based reinforcement learning', *Applied Sciences*, Vol. 13, No. 20, p.11321.

Guo, H. (2025) 'Piano harmony automatic adaptation system based on deep reinforcement learning', *Entertainment Computing*, Vol. 52, p.100706.

Jamshidi, F., Pike, G., Das, A. and Chapman, R. (2024) *Machine Learning Techniques in Automatic Music Transcription: A Systematic Survey*, arXiv preprint arXiv:2406.15249.

Ji, S., Yang, X., Luo, J. and Li, J. (2023) 'RL-chord: CLSTM-based melody harmonization using deep reinforcement learning', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 35, No. 8, pp.11128–11141.

Kamal, M.B., Khan, A.A., Khan, F.A., Shahid, M.M.A., Kamal, M.D. and Ali, M.J. (2022) 'An innovative approach utilizing binary-view transformer for speech recognition task', *Computers, Materials & Continua*, Vol. 72, No. 3, pp.5547–5562.

Large, E.W., Roman, I., Kim, J.C., Cannon, J., Pazdera, J.K., Trainor, L.J. et al. (2023) 'Dynamic models for musical rhythm perception and coordination', *Frontiers in Computational Neuroscience*, Vol. 17, p.1151895.

Latif, S., Cuayáhuitl, H., Pervez, F., Shamshad, F., Ali, H.S. and Cambria, E. (2023) 'A survey on deep reinforcement learning for audio-based applications', *Artificial Intelligence Review*, Vol. 56, No. 3, pp.2193–2240.

Lei, S. and Liu, H. (2022) 'Deep learning dual neural networks in the construction of learning models for online courses in piano education', *Computational Intelligence and Neuroscience*, Vol. 2022, No. 1, p.4408288.

Li, W. (2022) 'Analysis of piano performance characteristics by deep learning and artificial intelligence and its application in piano teaching', *Frontiers in Psychology*, Vol. 12, p.751406.

Liang, J. (2023) 'Harmonizing minds and machines: survey on transformative power of machine learning in music', *Frontiers in Neurorobotics*, Vol. 17, p.1267561.

Liang, Y. and Pan, F. (2023) 'Study of automatic piano transcription algorithms based on the polyphonic properties of piano audio', *IEIE Transactions on Smart Processing & Computing*, Vol. 12, No. 5, pp.412–418.

Liu, L. and Zhu, W. (2025) 'Research on restructuring the piano lesson teaching model in the context of artificial intelligence', *GBP Proceedings Series*, Vol. 5, pp.167–173.

Marták, L.S., Kelz, R. and Widmer, G. (2022) 'Balancing bias and performance in polyphonic piano transcription systems', *Frontiers in Signal Processing*, Vol. 2, p.975932.

Peter, SD. (2023) *Online Symbolic Music Alignment with Offline Reinforcement Learning*, arXiv preprint arXiv:2401.00466.

Phatnani, K.S. and Patil, H.A. (2024) 'Modeling musical expectancy via reinforcement learning and directed graphs', *Multimedia Tools and Applications*, Vol. 83, No. 10, pp.28523–28547.

Shang, R. (2022) 'A deep learning-enabled composition system based on piano score recognition', *Mobile Information Systems*, Vol. 2022, No. 1, p.9132697.

Wang, L., Zhao, Z., Liu, H., Pang, J., Qin, Y. and Wu, Q. (2024a) 'A review of intelligent music generation systems', *Neural Computing and Applications*, Vol. 36, No. 12, pp.6381–6401.

Wang, X., Ma, Z., Cao, L., Ran, D., Ji, M., Sun, K. et al. (2024b) 'A planar tracking strategy based on multiple-interpretable improved PPO algorithm with few-shot technique', *Scientific Reports*, Vol. 14, No. 1, p.3910.

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*, Vol. 30, pp.12967–12989.

Wei, W., Li, P., Yu, Y. and Li, W. (2022) *Hppnet: Modeling the Harmonic Structure and Pitch Invariance in Piano Transcription*, arXiv preprint arXiv:2208.14339.

Yaqoob, A., Yuan, Z. and Muntean, G.M. (2024) 'A UAV-centric improved soft actor-critic algorithm for qoe-focused aerial video streaming', *IEEE Transactions on Vehicular Technology*, Vol. 73, No. 9, pp.13498–13512.

You, W. (2024) 'Modeling method for classification of piano music style based on big data mining and machine learning', *IEIE Transactions on Smart Processing & Computing*, Vol. 13, No. 2, pp.129–139.

Zhai, Y. and Xu, C. (2022) 'The application of artificial intelligence-assisted computer on piano education', *Comput. Aided. Des. Appl*., Vol. 20, pp.157–167.

Zhang, M. (2025) 'Advancing deep learning for expressive music composition and performance modeling', *Scientific Reports*, Vol. 15, No. 1, p.28007.

Zhao, Z., Li, Q., Zhang, Z., Cummins, N., Wang, H., Tao, J. and Schuller, B.W. (2021) 'Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition', *Neural Networks*, Vol. 141, pp.52–60.