ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Problems and optimization of traditional Dulcimer teaching Effect and development System based on artificial intelligence application

Dongdong Ding[1,*]

[1] Communication University of ZhejiangMusic Academy 314500

SUMMARY: *In order to improve the ability of traditional dulcimer teaching effect evaluation, state recognition and optimization generation, this paper constructs a multi-modal analysis framework driven by artificial intelligence. Teaching audio, performance movements, classroom videos, and practice logs were collected from 96 learners in 16 teaching units to form a teaching dataset containing 18,240 synchronized samples. The model is composed of acoustic branch, action branch and classroom behavior branch. Time convolution, graph modeling and cross-modal attention mechanism are combined to jointly encode timespan stability, rhythm deviation, chord tapping posture and classroom interaction behavior, and output teaching scores and state labels. In the training phase, population search, AdamW and dynamic learning rate scheduling are used to complete the optimization, and the data set is divided by 7:2:1 on the NVIDIA RTX 4090 platform to complete the training and testing. The experimental results show that the proposed framework achieves 93.6% accuracy, 0.918 Macro-F1 and 24.6 ms reasoning delay, which provides a computable basis for the fine analysis and digital adaptation of traditional dulctic classroom.*

KEYWORDS: *Multimodal perception; Deep learning; Intelligent evaluation; Instructional Information System*

1 Introduction

1.1 Research background and question proposal

Under the background of the evolution of artificial intelligence, digital audio processing and educational computing technology, the traditional dulcimer teaching is shifting from experience teaching to data-supported analysis. The process of dulcpiano performance includes multi-source information such as pitch control, rhythm organization, chord strength, hand trajectory and classroom interaction. It is difficult to describe the formation process of teaching effect by auditory judgment alone, and it is also difficult to support the adaptation of different learning stages. Combining acoustic signal analysis, action recognition, time series modeling and instructional information system, performance, classroom feedback and learning process can be uniformly mapped into a computable space, which provides a data basis for instructional analysis.

Focusing on intelligent music teaching, teaching system modeling and algorithm support mechanism, the existing research has formed a relatively clear technical accumulation, and provides a reference path for the computational analysis of traditional dulctic teaching. Cui studied the application of piano teaching supported by augmented reality, indicating that

*13750815496@163.com
https://doi.org/10.65102/is2026218 Published: 30 April, 2026

intelligent interaction technology has entered the instrumental music training scene [1]. Chen proposed the design method of intelligent music teaching system based on virtual reality, which provides reference for the structure modeling of teaching system [2]. Yuan studied the implementation path of artificial intelligence assisted music teaching system and strengthened the perspective of system collaborative analysis [3]. Chen and Sun proposed the application framework of music education system supported by deep learning, which expanded the scope of teaching behavior modeling [4]. Song designed a remote piano teaching method integrating multi-head convolutional network and optimization mechanism, which reflects the evaluation ability driven by algorithm [5]. Wang built a platform system for vocal music teaching, which verified the feasibility of intelligent teaching platform [6]. Chen studied the application of big data and fuzzy decision support in personalized music teaching, which enriched the expression of teaching decision [7]. Another Wang proposed a personalized recommendation system for zither based on deep learning, indicating that the teaching of folk instrumental music has entered the stage of algorithm support [8]. Su and Wang applied reinforcement learning to adjust music teaching strategies and enhanced the dynamic adaptation ability [9]. Fang studied the virtual learning scheme for intelligent music education system, which extended the deployment idea [10]. These studies have promoted the development of intelligent music teaching from the aspects of interaction forms, model design, recommendation mechanisms and platform construction. However, most of the existing results are oriented to piano, vocal music or general music education scenarios.

The existing results provide a method for intelligent teaching of dulcimer, but the timbre granularity, the discreteness of clinging action and the coupling characteristics of classroom process in dulcimer scene still need to be modeled. Based on this, this paper focuses on the evaluation of traditional dulcimer teaching effect and system adaptation, and constructs a computational link between multimodal data organization, feature extraction, model training and result output, so as to form a technical framework that takes into account both teaching application and information system implementation.

1.2 Review of related research and technical gap

The research on intelligent traditional instrumental music teaching is shifting from single experience judgment to multi-modal calculation analysis. Jiang studied the fuzzy clustering method in distance music education, which provided the basis for resource stratification [11]. Chen studied the piano training of augmented reality and Internet of things, and showed that the perception terminal and teaching feedback can form a linkage relationship [12]. Huang and Ding proposed an intelligent piano teaching path, which realized the connection between teaching process and algorithm support [13]. Ji, Wang, and Wang studied the application of deep reinforcement learning in piano skill training to extend the computational space for teaching strategy generation [14]. These results show that artificial intelligence is no longer limited to auxiliary display, but gradually enters the core links of teaching process analysis, training feedback regulation, and learning state modeling.

In terms of performance evaluation and behavior recognition, Ghatas, Fayek and Hadhoud proposed a hybrid deep learning method for musical difficulty estimation of piano symbols, and proved that complex performance processes can be mapped into stable feature representations [15]. Ramoneda, Jeong and Eremenko et al. studied the difficulty classification of music scores under multi-dimensional piano performance fusion, which provided reference for the quantification of teaching effect [16]. Enkhbat, Shih and Cheewaprakobkit proposed human action recognition and note recognition methods based on STA-GCN to make the correspondence between action trajectories and performance results clearer [17]. However,

there is still a lack of a unified multi-modal modeling framework among chord percussion path, timbral dispersion, rhythm organization and classroom behavior in dulcimer scene. When existing methods are directly transferred to traditional dulcimer teaching, there are still obvious limitations in feature adaptation and result interpretation.

On the other hand, the interpretation of teaching effect also needs to combine emotional understanding and perceptual differences. Koh, Cheuk and Heung et al. constructed a music dataset containing emotion ratings and rater portraits, which provided support for subjective perception modeling [18]. Medina, Beltran, and Baldassarri studied music emotion classification based on neural networks and verified the role of emotion labels in music understanding [19]. Louro, Redinho and Malheiro et al. compared a variety of deep learning music emotion recognition methods and revealed the influence of model structure differences on recognition performance [20]. Jiang, Zhang and Lin et al. systematically reviewed the research on music emotion recognition based on deep learning and pointed out that multi-modal fusion and feature expression are still the main promotion path in this direction [21]. The existing results provide a method reference for the computational analysis of traditional dulctic teaching, but the integrated information system for teaching effect evaluation, state recognition, optimization generation and application adaptation still needs to be further constructed.

2 Data modeling and basic system construction of traditional dulcimer teaching scene

2.1 Feature analysis of traditional Dulcimer teaching process

The structural analysis of traditional dulctic teaching process is the premise of teaching effect calculation and system modeling. Compared with the standard general music training scene, dulcpiano teaching includes the process characteristics of acoustic feedback, chord percussion, spectral surface understanding and classroom interaction, and different learning stages show differences in speed control, timbral stability and bimanual collaboration. Only when these process information is extracted accurately, the subsequent evaluation model, recognition module and optimization strategy can obtain reliable input.

At the level of acoustic performance, dulcpiano performance emphasizes the harmonious relationship between clear grain, balanced strength and rhythm fluctuation. In the teaching process, the same phrase will form different envelope changes and frequency band energy distribution in the state of elicitation, legato and legato. To describe this difference, short-term energy sequence, spectral centroid trajectory and onset sharpness are jointly encoded to characterize timbre control and rhythm execution changes along the time axis.

At the level of action execution, dulccipital teaching has the characteristics of alternating percussion of both hands, and the range of wrist rotation, percussion Angle and position of the hammer will directly affect the sound formation. In order to maintain the consistency between movement information and acoustic output, this paper uses the key point trajectory sequence and local displacement statistics to describe the performance posture, and performs the timing alignment of the left-hand beat correspondence, so as to capture the movement stability and coordination pattern in the training phase.

At the classroom process level, teacher demonstration, student imitation, segmented correction and practice constitute the teaching promotion link. This paper introduces three indicators: practice round density, error correction response interval and segment completion rate to discretize the classroom process, so that teaching behavior, performance and feedback results can enter a unified data expression space.

The process tensor constructed based on the above features will be used as the input of the

subsequent intelligent evaluation model. In order to enhance the sample coverage, the representative clips at three levels of beginner, advanced and improved were added, and the balanced sampling was carried out with different speed, mode and rhythm types of tracks, so that the proportion of low-frequency teaching situations in the data set was increased from 12% to 24%, so as to improve the adaptation ability of the model to complex teaching scenes.

2.2 Collection and standardization of teaching audio, performance movements and classroom process data

In order to build a basic data system suitable for the evaluation of traditional dulctic teaching effect and system adaptation analysis, this paper establishes a synchronous collection and standardized processing process around three types of information: teaching audio, performance action and classroom process, and provides consistent data input for subsequent model training, state recognition and optimization generation. The data collection covered three learning levels of beginner, advanced and improved, including solo clips, sentence exercise clips and teacher demonstration clips of 96 learners in 16 teaching units, forming a total of 18,240 groups of synchronous samples. The audio signal is uniformly saved with 48 kHz sampling rate and 24 bit quantization precision, and segmented according to typical teaching fragments such as initiation, legato, round bamboo and fast door, so as to ensure the complete retention of fine-grained performance behavior in dulc Qing classroom.

For instructional audio processing, the study performed environmental noise suppression, loudness alignment, and segment boundary correction on the original recordings, and then generated time-frequency representations by short-time Fourier transform and Mel-filter bank. In order to weaken the amplitude offset caused by different classroom conditions and the difference of percussion strength, this paper implements the normalized modeling of local frequency band energy, and its expression is as follows.

$$M_{t,f} = \log\left(1 + \frac{|X_{t,f}|^2}{\mu_f + \varepsilon}\right) \cdot \omega_t \tag{1}$$

Here, $M_{t,f}$ represents the normalized spectral energy of frequency band f in frame t, $X_{t,f}$ represents the short-time spectral coefficient, μ_f represents the local band mean, ε represents the stabilization term, and ω_t represents the onset enhancement weight. This formula is used to retain teaching-related acoustic information such as chord striking transient, overtone diffusion and residual vibration attenuation. This processing makes the spectral features between different practice rounds comparable, and also facilitates the subsequent network to extract indicators such as rhythm stability, timbre clarity, and strength consistency.

In terms of performance action and classroom process processing, the system adopts dual-camera video, wrist inertial sensor and classroom log to synchronously record chord stroke trajectory, wrist rotation Angle, mallet position, teacher correction time and student repetition rounds, and the video frame rate is unified as 60 frames per second. The key point detection module extracts the positions of shoulder, elbow, wrist and the end of the piano and bamboo, and then combines the classroom events to complete the timing alignment. In order to realize the joint coding of action flow, rhythm flow and classroom flow, the key point displacement, beat response intensity and classroom event coding are mapped into a unified state space, and their fusion expression is as follows.

$$Z_t = \alpha \sum_{i=1}^{K} A_{t,i}\, \Delta p_{t,i} + \beta r_t + \gamma c_t \tag{2}$$

Here, Z_t represents the joint state vector At time t,$A_{t,i}$ represents the visibility weight of key points, $\Delta p_{t,i}$ represents the displacement of adjacent frames, r_t represents the beat response intensity, c_t represents the classroom event coding, and α, β, γ represent the fusion coefficients. This formula maps action stability, beat execution and teaching advancement into a unified time sequence space, which can directly serve the parallel loading and state discrimination of subsequent multimodal networks.

After collection, all samples are processed by denoising, resampling, missing imputation, label alignment and tensor encapsulation in sequence, and organized as standard input units in the form of audio matrix, action sequence and classroom event vector, which has strong interface sharing ability with cross-modules. This process ensures the consistent expression of acoustic layer, action layer and teaching layer, and also provides a stable data basis for the calculation of traditional dulcine teaching effect, system state recognition and application adaptation analysis.

2.3 Teaching effect evaluation label system and database structure design

In order to support the training of traditional dulcimer teaching effect evaluation model and the deployment of teaching information system, this paper constructs a tag system and database structure with extensible, searchable and computable expression ability. Common teaching data records mostly stay at the level of tracks, learners and classes, which are difficult to carry the correlation expression between audio performance, action state and classroom feedback. Based on this, this paper adopted a multi-dimensional and multi-granularity annotation method and a hierarchical database architecture to realize the collaborative organization of performance features, teaching results and system states.

From the perspective of label design, this paper divides the teaching sample into three dimensions: performance layer, process layer and feedback layer. The performance layer describes the rhythm stability, timbre clarity and strength consistency, the process layer records the chord striking action, clause completion and correction response, and the feedback layer describes the evaluation level, adaptation suggestions and system status. To facilitate subsequent calls, the core label structure is shown in Table 1.

Table 1: Label system for teaching effect evaluation

Label Dimension	Main Fields	Functional Orientation
Performance Layer	Rhythm Stability, Timbre Clarity, Dynamics Consistency	Describes the quality of performance output
Process Layer	Striking Trajectory, Phrase Completion Degree, Error-Correction Response Time	Represents the progression of the teaching process
Feedback Layer	Evaluation Level, Adaptive Suggestions, System State Labels	Supports result output and optimization generation

In the aspect of database structure design, this paper adopts the hybrid organization method of relational database and document database. The relationship layer established the sample primary key, course number, learner number and label index based on PostgreSQL to ensure the stability of statistical query, version control and authority management. The document layer encapsulates the audio path, spectrum matrix, action sequence, classroom event stream and label vector of a single sample in JSON, which supports the parallel reading of multi-modal tensors. This structure is not only suitable for batch construction in the deep learning framework, but also convenient for the front-end teaching platform to call evaluation results, playback clips and generate personalized records.

In order to ensure label consistency and data security, the system sets hash verification, field constraints and semantic mapping rules in the writing phase, and performs automatic screening of abnormal samples. Similar tags were merged through the mapping table, cross-class records were associated through timestamp and learner identification, and duplicate segments were removed through fingerprint comparison. At the same time, the label evolution log is kept to facilitate the subsequent error backtracking and structure revision. Improve database maintenance transparency. The overall database adopts modular interface design, so that label update, sample addition and model retraining can continue to run on the same information base, so as to provide stable support for traditional dulctic teaching effect calculation, system state recognition and application adaptation analysis.

3 Construction of problem identification model of traditional Dulcimer teaching effect evaluation and development system

3.1 Design of multimodal teaching effect evaluation model

The traditional dulctic teaching effect evaluation relies on chord tapping action and classroom advancement information at the same time, and a single network is difficult to take into account acoustic changes, hand movement differences and teaching rhythm evolution. Based on this, this paper constructs a multimodal assessment model, which consists of an input mapping layer, an acoustic branch, an action branch, a classroom branch, a fusion layer and a double-ended output layer. The model input includes spectrum tensor, key point sequence and classroom event coding. The three types of data are aligned with the time axis and the dimension projection is completed to ensure that different modalities maintain correspondence in the same segment. In order to achieve a unified expression, this paper uses a shared mapping function:

$$h_t = \sigma(W_a a_t + W_m m_t + W_c c_t + b_h) \tag{3}$$

Here, h_t represents the joint input vector at time t; W_a, W_m and W_c represent the mapping matrices of audio, action and classroom events respectively; a_t, m_t and c_t represent the three types of original inputs; b_h represents the bias term; and σ represents the activation function. This equation is used to weaken the scale differences of heterogeneous modalities.

The acoustic branch uses the series structure of residual convolution and dilated convolution to locally perceive the dulcharp chord transient, residual vibration and frequency band migration. The convolutional output goes into the short-range aggregation unit after the channel recalibination to enhance the ability to distinguish leghatto, turban and weak onset segments. Its status update is written as follows:

$$S_l = \Phi(S_{l-1}) + \eta D_r(S_{l-1}) \tag{4}$$

Here, S_l represents the l layer acoustic feature map, Φ represents the convolution transform, D_r Represents the dilated convolution with dilation rate r, and η represents the residual retention coefficient. This formula preserves local texture and cross-frame energy diffusion information.

The action branch constructs a spatio-temporal graph with the key points of the skeleton and the end points of the harp and bamboo as nodes, and the edge weight is determined by the joint displacement and the string striking sequence. The classroom branch encodes the teacher demonstration, pause correction, repeat exercise and completion marks as an event sequence,

and the gating unit captures the rhythm change of teaching progress. To form joint constraints, this paper combines the two types of hidden states as follows.

$$g_t = P[u_t; v_t] \odot \tanh(u_t + v_t) \tag{5}$$

Here, g_t represents the teaching behavior representation, u_t represents the action graph convolution output, v_t represents the hidden state of classroom events, P represents the projection matrix, and $\odot$ represents element-wise modulation. This equation binds chord stability and classroom response intensity to a unified space.

In the fusion layer, the cross-modal attention mechanism is used to calculate the contribution of each modality to the segment, and the results are sent to the effect evaluation end and the state recognition end respectively. The former outputs the completion quality score, and the latter gives the state probabilities of normal, need adjustment and key intervention. The final prediction is as follows:

$$y = \text{softmax}(W_o F_t + b_o) \tag{6}$$

Here, y represents the output vector, F_t represents the fusion feature, W_o represents the output mapping matrix, and b_o represents the bias term. This formula unifies the continuous rating and discrete state into the same decision space, which provides a stable basis and support for subsequent optimization generation.

The above model completed the collaborative modeling of acoustic information, motion information and classroom behavior information in a unified input space, so that the performance of dulcimer teaching and the process of teaching promotion could be described at the same time. The structure not only retains fine-grained performance differences, but also enhances the consistency of teaching state expression, which provides stable support for subsequent system identification and optimization generation.

3.2 Acoustic features, motion features and classroom behavior feature extraction methods

In the traditional dulcute teaching scene, acoustic features, motion features and classroom behavior characteristics jointly determine the expression boundary of teaching effect evaluation. In order to ensure that the multimodal input not only retains the performance details, but also has a unified computing interface, this paper constructs a hierarchical feature extraction process, which completes signal transformation, trajectory coding, classroom event parsing and tensor encapsulation in turn.

The overall process is shown in Fig. 1. The audio end extracts the acoustic features such as MEL spectrum, spectral centroid and spectral flow, the video end extracts the key point trajectory and action change information, and the classroom log end completes the coding of teaching events. After time alignment and tensor encapsulation, the three types of features form a unified input, and are fed into the subsequent teaching effect evaluation model for effect scoring, state recognition and optimization suggestions generation.

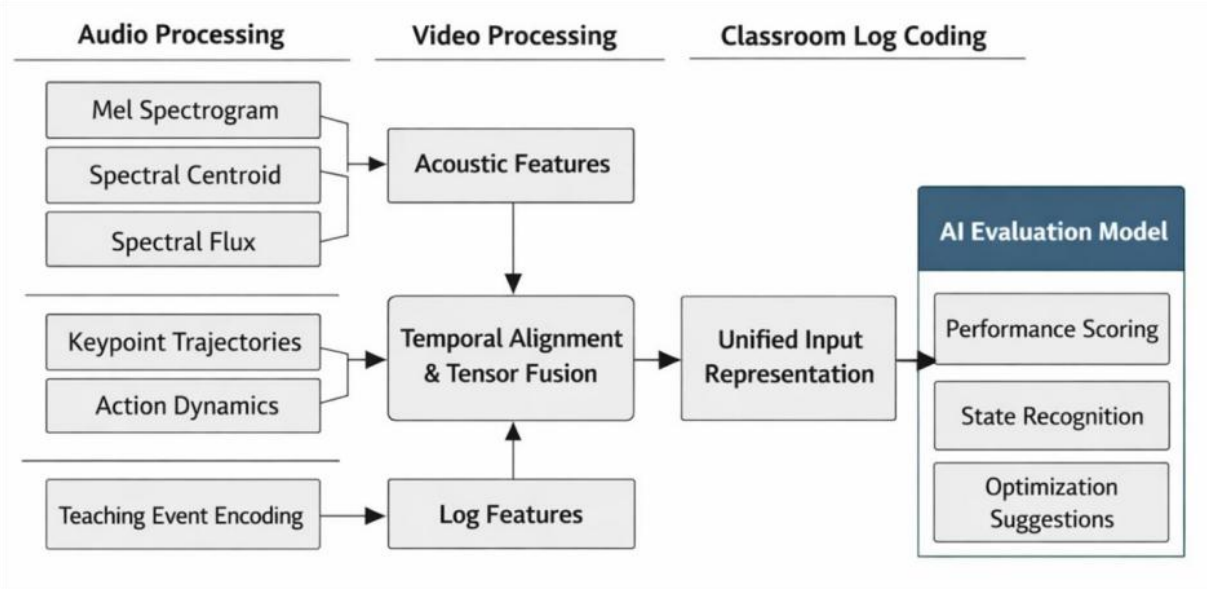

Figure 1: Flowchart of multi-source feature extraction and unified coding for traditional dulcute teaching

Acoustic feature extraction focuses on describing dulcimer chord transient, residual vibration attenuation, strength fluctuation and beat balance. In this paper, a joint representation of MEL spectrum, spectral centroid, spectral flow and transient energy is used, and adaptive band compression is used to weaken the interference of high-frequency noise on teaching clips. In order to characterize the degree of time-frequency energy aggregation, the local acoustic description formula is defined as follows.

$$A_t = LN(M_t + \lambda_1 c_t 1 + \lambda_2 f_t 1) \tag{7}$$

Here, A_t represents the acoustic representation at time t, M_t represents the MEL spectral vector, c_t represents the spectral centroid, f_t represents the spectral flow, λ_1 and λ_2 represent the weight coefficients, 1 represents the spread vector, and LN represents the layer normalization. This formula can simultaneously retain the timbre clarity and rhythm fluctuation information.

The motion feature extraction is carried out around the two-handed string stroke trajectory, the wrist lift amplitude and the mallet drop stability. Firstly, the system extracted the key points of shoulder, elbow, wrist and the end of the harp and bamboo in each frame, and then calculated the displacement of adjacent frames and the change of chord Angle. In order to describe the action coherence, the posture evolution expression is constructed as follows.

$$B_t = \rho B_{t-1} + (1-\rho)\,[\Delta p_t; \Delta\theta_t; v_t] \tag{8}$$

Here, B_t represents the action state vector, Δp_t represents the keypoint displacement, $\Delta\theta_t$ represents the wrist rotation Angle change, v_t represents the endpoint velocity, and ρ represents the smoothing coefficient. This formula can reflect the action amplitude, direction consistency and beat synchronization degree.

Classroom behavior features are used to describe teaching promotion information such as teacher demonstration, pause correction, clause repetition and completion confirmation. After embedding, the event sequence enters the temporal convolution unit, and the correction nodes and exercise-intensive areas are highlighted by gated weights. The encoding is written as

follows:

$$C_t = \mathrm{GELU}(W_c[e_t; q_t; d_t] + b_c) \tag{9}$$

Here, C_t represents classroom behavior vector, e_t represents event embedding, q_t represents repetition intensity, d_t represents correction interval, W_c represents mapping matrix, and b_c represents bias term. Finally, the acoustic, movement and classroom behavior features were spliced into a unified input tensor according to the time slice, which made the traditional dulctic teaching effect evaluation have a stable multi-source expression basis and enhanced system adaptability.

After the above extraction and coding, acoustic features, action features and classroom behavior features are organized into a unified multimodal input unit. In this way, the timbre change, chord strike control and classroom advancement in dulccipital teaching can enter the same computational space, which provides a consistent data basis for subsequent effect evaluation, state recognition and result generation.

3.3 Artificial intelligence algorithm design and model training strategy

In the traditional dulcute teaching effect evaluation task, the model training strategy not only affects the parameter convergence speed, but also directly relates to the stability of teaching state discrimination and result generation. In this paper, a training framework of "pre-training initialization-phased update-dynamic constraint correction" is adopted to gradually converge the acoustic branch, action branch and classroom behavior branch within a unified loss space. At the beginning of training, the cross-modal alignment layer is frozen, and only the local parameters of each branch are updated to maintain the independence of the feature boundaries of different modalities. The fusion layer was thawed in the mid-term to enhance cross-modal collaboration. At the later stage, the state constraint term is introduced to keep the output results consistent with the teaching scene. In order to control the gradient fluctuation under multi-task learning, this paper constructs a joint loss function with adaptive weights

$$L = \lambda_1 L_{\mathrm{score}} + \lambda_2 L_{\mathrm{state}} + \lambda_3 L_{\mathrm{cons}} + \lambda_4 \|\Theta\|_2^2 \tag{10}$$

Here, L represents the total loss, L_{score} represents the teaching effect scoring error, L_{state} represents the teaching state classification error, L_{cons} represents the cross-modal consistency constraint, Θ represents all trainable parameters, and λ_1 to λ_4 represent the dynamic weights. This formula is used to unify the continuous score, discrete state and structural regularity into the same optimization objective, avoiding the deviation of the model on a single task.

In the parameter update stage, the combination optimization strategy of AdamW and Lookahead is used to balance the rapid descent in the early stage and the stable approximation in the later stage. Its updated form is written as follows:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

$$\Theta_{t+1} = \Theta_t - \eta \frac{m_t}{\sqrt{v_t} + \varepsilon} - \eta \omega \Theta_t \tag{11}$$

where g_t represents the current batch gradient, m_t and v_t represent the first and second moment estimates, η represents the learning rate, ω represents the weight decay coefficient, and ε represents the numerical stability term. This formula not only retains the advantage of

adaptive step size, but also suppresses the oscillation of deep network in high-dimensional parameter space.

In order to improve the adaptation ability of structural configuration, a population search mechanism is introduced before training to jointly optimize the size of convolution kernel, the depth of graph convolution, the dimension of hidden units and the number of attention heads. Its fitness function is defined as follows.

$$F = \alpha Acc_{val} + \beta F1_{macro} + \gamma\left(1 - \frac{T_{inf}}{T_{max}}\right) + \delta\left(1 - \frac{M_{use}}{M_{max}}\right) \tag{12}$$

where Acc_{val} represents the validation set accuracy, $F1_{macro}$ represents the macro average F1 value, T_{inf} represents the inference delay, M_{use} represents the video memory footprint, and α, β, γ, δ represent the weights. This formula takes performance and deployment cost into the search criteria at the same time, so that the model training results are more suitable for the teaching platform.

In terms of training scheduling, this paper adopts a joint mechanism of cosine annealing and warm start to keep the learning rate controllable in different stages:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\frac{\pi \cdot \mathrm{mod}(t, T_r)}{T_r}\right) \tag{13}$$

Here, η_t represents the learning rate at round t, η_{max} and η_{min} represent the maximum and minimum learning rates, respectively, and T_r represents the current cycle length. This strategy can alleviate the late training stagnation and enhance the tolerance of the model to small batch perturbations.

On the whole, the algorithm design of this paper is not simply the pursuit of training speed, but around the multi-source information coupling characteristics in the dulcimer teaching scene, a training mechanism that takes into account accuracy, stability and deployment adaptability is constructed. After phased update, joint loss constraint, structure optimization and dynamic scheduling, the model can stably output teaching effect scoring and state discrimination results, and provide a reliable parameter basis for subsequent teaching adaptation and optimization results generation.

3.4 Teaching effect output structure and system problem identification mechanism

In order to achieve a stable mapping from fusion features to the expression of teaching results, this paper constructs a three-level mechanism of "effect scoreing-state identification-suggestion triggering" at the output end. The structure can not only retain continuous quantitative results, but also generate discrete category labels, so that the system can simultaneously serve classroom feedback, learning record update and teaching adaptation call. Firstly, the feature vector output by the fusion layer enters the score mapping unit, and performs a comprehensive regression on the rhythm stability, timeliness control, action coordination and classroom completion. Then it entered the state recognition unit to determine the teaching state to which the current segment belonged. Finally, the corresponding feedback results are triggered according to the confidence and rule constraints.

In the stage of teaching effect scoring, this paper uses weighted fusion regression to generate comprehensive scores:

$$s = \sum_{i=1}^{n} \pi_i r_i, \quad \pi_i = \frac{\exp(u_i)}{\sum_{j=1}^{n} \exp(u_j)} \tag{14}$$

Here, s represents the comprehensive teaching effect score, r_i represents the regression output of the i evaluation subitem, π_i represents the corresponding weight, and u_i represents the scoring weight generated by the attention mechanism. This formula enables different evaluation sub-items to automatically adjust their contribution proportion according to the characteristics of the segment, thereby avoiding information compression caused by fixed weighting.

In the state recognition stage, the system sends the fusion features to the entry control classifier to make probability discrimination of three kinds of states: "stable completion, local deviation, and need key intervention". The expression is:

$$p_t = \mathrm{softmax}(W_p F_t + b_g + G_t \odot U_g F_t) \tag{15}$$

Here, p_t represents the state probability vector, F_t represents the fusion feature, W_p represents the classification mapping matrix, b_g represents the bias term, G_t represents the gating coefficient, U_g represents the state modulation matrix, and $\odot$ represents the element-wise action. This formula strengthened the action imbalance, beat offset and classroom response hysteresis signals in the key segments through the gating gain, so that the state output was closer to the actual teaching process.

In order to enhance the reliability of the system output, a confidence consistency constraint is added between the scoring and the classification results. Only when the continuous score and the discrete state meet the same direction of change, the system will write the final discrimination result. The mechanism is defined as follows.

$$q_t = \sigma(w_s \cdot s + w_p \cdot \max(p_t) - \tau) \tag{16}$$

Here, q_t represents the confidence of the result, w_s and wp represent the weight of the scoring and classification items respectively, $\max(p_t)$ represents the highest state probability, τ represents the trigger threshold, and σ represents the Sigmoid function. When q_t is higher than the set threshold, the current output will be retained and enter the subsequent feedback module.

At the mechanism implementation level, the scoring results are written into the learning record table, the status labels are synchronously written into the teaching monitoring table, and the trigger tags are sent to the proposal generation module. In this way, the system no longer only gives a single value, but forms a joint output structure of "score-state-trigger", so that the classroom segments, training records and system calls maintain a consistent data interface. Finally, teaching effect output and status recognition can be expressed in a unified framework, which provides a clear, stable and traceable result basis for subsequent optimization result generation, application adaptation and visual presentation.

3.5 Optimization result generation and teaching application adaptation design

After obtaining the teaching effect scoring and status recognition results, the system needs to further generate executable optimization results, and complete the application adaptation according to the classroom terminal, training mode and learning level. To this end, this paper

constructs a processing link of "deviation analysis-result production-scene adaptation", so that the model output can be directly transformed into callable teaching aid results. The optimization results are not simply unified tips, but differentiated suggestions for rhythm control, timbre adjustment, chord striking action and classroom promotion rhythm, which are sent to the teacher end and the learning end through rule mapping and interface encapsulation.

In the deviation analysis stage, the system first calculates the distance between the current sample and the target representation vector, and forms the optimization direction according to the deviation degree of each dimension. Its expression is:

$$d_t = R(y_t^* - \hat{y}_t) + (I - R)\bar{e}_t \tag{17}$$

where d_t represents the optimization direction vector, y_t^* represents the target performance vector, $\hat{y}_t$ represents the current output vector, R represents the dimension selection matrix, I represents the identity matrix, $\bar{e}_t$ represents the mean of the historical residuals of the classroom. In this formula, the immediate offset and the phased cumulative offset are jointly considered, so that the generated results reflect not only the current segment state, but also the continuous training trajectory.

In the result generation phase, the system calculates the priority of the suggested items based on the optimization direction vector, and outputs the recommended order of rhythm, movement, strength and clause exercises in a ranked manner. The scoring function is written as follows:

$$r_k = \xi_1 g_k(d_t) + \xi_2 h_k(s_t) + \xi_3 m_k(h_t) \tag{18}$$

Here, r_k represents the priority score of the k proposal, $g_k(d_t)$ represents the deviation response strength triggered by the optimization direction, $h_k(s_t)$ represents the adaptation degree of the current teaching state to the proposal item, $m_k(h_t)$ represents the benefit estimate in the historical training records, and ξ_1 to ξ_3 represent the weighting coefficients. This formula can unify the immediate judgment, state matching and historical experience into the same ranking framework, and avoid the disconnection of the recommendation results.

In the application adaptation stage, the system also needs to adjust the output form according to the device capabilities and classroom scenarios. In the face of three kinds of application entries, teacher terminal, student terminal and mobile terminal, this paper uses constraint mapping objective function to control the compression and presentation granularity of results:

$$J = \arg\min_{\Omega}\{L_c(\Omega) + \lambda_1 L_a(\Omega) + \lambda_2 L_d(\Omega)\} \tag{19}$$

Here, Ω represents the set of adaptation configurations, L_c represents the content complexity cost, L_a represents the application scenario matching cost, L_d represents the device resource consumption cost, and λ_1 and λ_2 represent the balance coefficients. This formula is used to strike a balance between feedback integrity, interface readability and terminal load, so that the same model output can remain stable and available in different teaching scenarios.

Through the above mechanism, the system can convert the model output into structural optimization results, and automatically adjust the presentation mode and call interface according to the application environment. In this way, the result feedback in traditional dulcute teaching no longer stays at the static evaluation layer, but forms a continuous output link for classroom use, after-class training and mobile learning, which provides unified support for teaching adaptation, record update and system expansion.

4 Experiment and result analysis

4.1 Experimental environment and parameter setting

In order to verify the traditional dulcimer teaching effect evaluation and system adaptation framework, this paper completed the experimental deployment under the consistent conditions of hardware, software and data. The experimental platform uses Ubuntu 22.04 operating system, Intel Core i9-13900K processor, NVIDIA RTX 4090 graphics card, and 128 GB memory. The development environment is built on Python 3.10, PyTorch 2.2 and CUDA 12.1, and connected to PostgreSQL and JSON interface modules. The dataset covers 96 learners, 16 teaching units, and 18240 sets of synchronized samples, which are divided into training, validation, and test sets by 7:2:1.

As shown in Table 2, the parameters are set around the input specification, training schedule, and evaluation manner.

Table 2: Experimental environment and parameter Settings

Item	Configuration
Operating System	Ubuntu 22.04
Development Framework	Python 3.10, PyTorch 2.2, CUDA 12.1
Hardware Platform	i9-13900K, RTX 4090, 128 GB RAM
Input Specifications	Audio 512×128, Video 60 fps, Event-Synchronized Encoding
Optimizer	AdamW + Lookahead
Initial Learning Rate	1×10-4
Batch Size	16
Number of Training Epochs	120
Early Stopping Strategy	Training stops if no improvement is observed on the validation set for 12 consecutive epochs
Evaluation Metrics	Accuracy, Macro-F1, Inference Latency

In terms of parameters, the input of the audio branch is a 512×128 spectrum tensor, the action branch receives a 60 fps keypoint sequence, and the classroom behavior branch uses an event embedding of length 64. AdamW and Lookahead combination optimizer was used in the training phase, the initial learning rate was set to 1×10-4, the minimum decay was set to 1×10-6, the batch size was set to 16, and the number of training rounds was 120. In order to suppress overfitting, the Dropout of the convolution branch is set to 0.3, and the Dropout of the temporal branch is set to 0.5. The weight attenuation and label smoothing mechanism are added to the fusion layer, and the gradient clipping threshold is set to 1.0. All models use the same data segmentation, preprocessing process and evaluation index, which provides the basis for subsequent test and analysis.

4.2 Analysis of test results of traditional Dulcute teaching effect

In order to verify the effectiveness of the constructed model in the traditional dulcute teaching scene, this paper carried out tests from three levels of teaching effect scoring accuracy, state discrimination consistency and multimodal response stability. The test set contains 1824 groups of samples, covering three levels: beginner, advanced and improved, and keeping the proportion of the three types of segments: solo, clause exercise and teacher demonstration. The results show that the Accuracy of the model on the overall test set is 93.6%, Macro-F1 is 0.918, and

the average inference delay is 24.6 ms. It shows that the multi-modal evaluation framework can meet the requirements of real-time analysis in classroom feedback while maintaining high recognition accuracy. The correlation coefficients between the four sub-indexes of rhythm stability, timbral clarity, movement coordination and classroom completion and the manual review results were 0.914, 0.927, 0.903 and 0.896, respectively, indicating that the model output was highly consistent with the actual teaching judgment.

As shown in Fig. 2, this paper uses confusion matrix to show the discrimination results of three types of teaching states. In the stable completion samples, 612 groups were correctly identified, accounting for 95.6% of the samples. In the local offset samples, 544 groups were correctly identified, and the recognition rate was 92.2%. 468 groups of key intervention samples were accurately classified, and the recognition rate was 91.8%. The off-diagonal regions were mainly concentrated between "local offset and key intervention", indicating that some high-intensity correction segments had similar characteristics in action trajectories and classroom behaviors, which also reflected that the model had strong discrimination ability for boundary states, but still kept prudent discrimination for transitional samples.

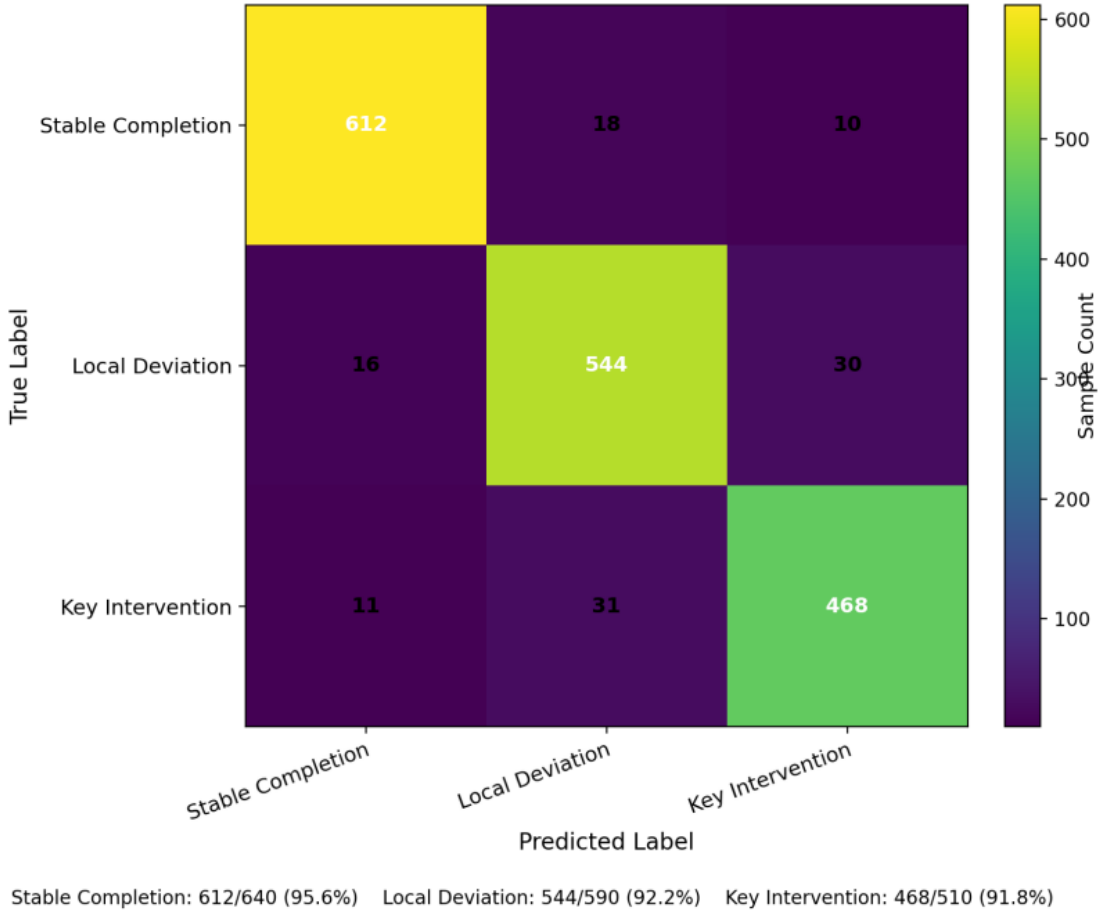

Figure 2: Confusion matrix for traditional dulcimer teaching state recognition

In the multi-dimensional teaching effect analysis, this paper further uses the radar chart to compare the output distribution of different learning levels. As shown in Fig. 3,the average scores of rhythm stability, movement coordination and classroom completion at the beginner level were 81.4, 78.9 and 80.6, respectively. The corresponding scores of the advanced level have been increased to 88.7, 86.2, and 87.4. The higher level reaches 93.1, 91.8 and 92.6. The

differences of the three types of samples in the timbre clarity dimension were relatively convergence, which were 84.8, 89.1 and 92.3, respectively, indicating that the improvement of the motion and rhythm dimensions was more obvious with the increase of training depth, and the timbre control showed more stable accumulation characteristics in the middle and late stages. The results are consistent with the training rule of "first stabilize the rhythm and movement, then refine the timbre" in dulcimer teaching.

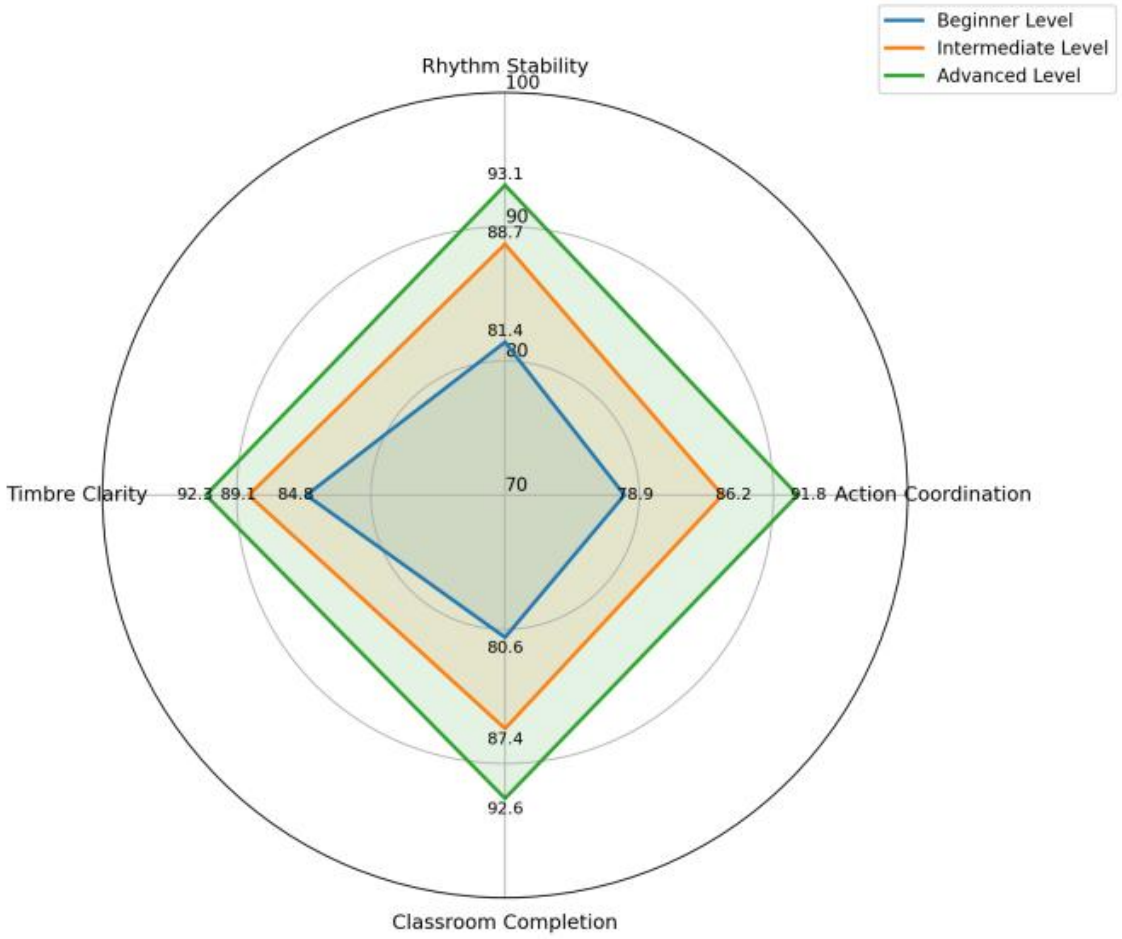

Figure 3: Multidimensional output radar chart of teaching effect at different learning levels

The comprehensive test results show that the constructed model is not only superior to the single-modal baseline in terms of overall accuracy, but also can more stably depict the rhythm, timbre, movement and classroom advancement information in traditional dulcimer teaching, which provides a reliable basis for subsequent system problem identification and optimization effect analysis. In the test, the standard deviation of each batch is lower than 0.04, indicating that the model maintains good output stability and result consistency between different classes and learning levels.

4.3 Problem identification and optimization effect analysis of traditional Dulcimer teaching system

In order to test the effect of problem recognition and optimization of the system in the traditional dulcimer teaching scene, this paper further carries out closed-loop verification on the basis of the test set. The verification object is still 1824 groups of segments, but three additional rounds of optimization iteration records are introduced to compare the performance changes of the

system after recognition, feedback and adjustment. The recognition end focuses on four systematic representations such as rhythm imbalance, chord deviation, timbre laxity and classroom response hysteresis, while the optimization end tracks the convergence speed, state decline proportion and suggestion adoption strength after feedback. The results show that the overall accuracy of system problem identification reaches 92.8%, and the macro-average F1 value is 0.907, which is 3.9%and 4.4% higher than that without introducing the classroom behavior branch, indicating that the multimodal link can more completely retain the corresponding relationship between performance and teaching progress in traditional teaching analysis of Yang Qing.

As shown in Fig. 4, this paper uses a two-dimensional projected clustering scatter plot to show the distribution differences of the three types of teaching phases in the problem feature space. The beginner, advanced and improving stages form relatively independent clustering regions, in which the beginner stage is mainly distributed in the lower right region, the advanced stage is concentrated in the upper region, and the improving stage is located in the lower left region. The average inter-class distance in the figure is 0.57, which indicates that different learning levels have good discrimination in the characteristics of rhythm control, chord stability, timbre state and classroom response. The retention rate of variance after two-dimensional projection reaches 0.98, which indicates that the current planar distribution can reflect the information structure of the original 4-D problem features more completely. The sample distribution is more concentrated in the advanced stage, and the dispersion degree is relatively higher in the beginner stage and the improvement stage.

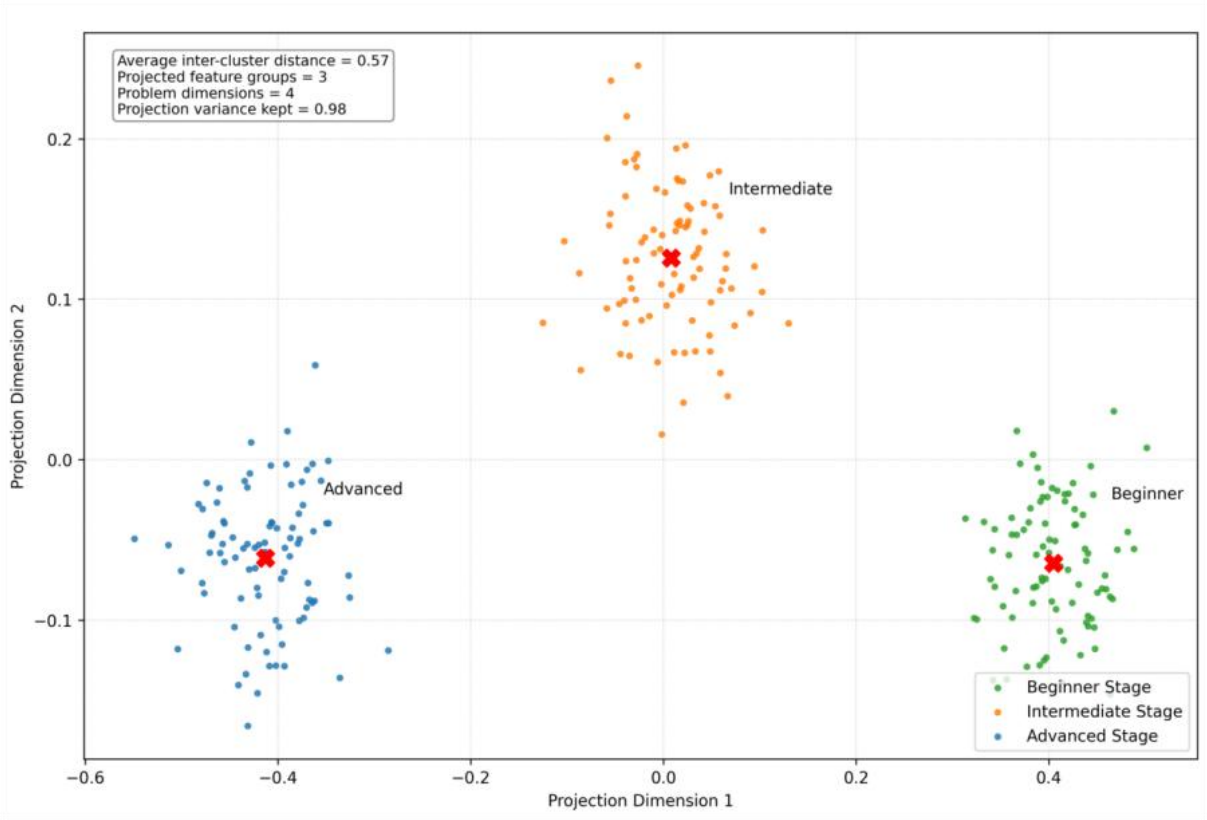

Figure 4: Scatter plot of problem distribution in traditional Dulcimer teaching system

The optimization results are further analyzed by the state transition ratio. As shown in Fig. 5, the system shows a clear direction of regulation for different initial state samples after proposal generation. Among the samples from "key intervention", 61.3% turned into "local

deviation", 18.6% further fell back to "stable completion", and the proportion of "key intervention" remained at 20.1%, indicating that high-risk segments had been significantly alleviated after one optimization. Among the samples from "local deviation", 57.8% entered "stable completion" after two rounds of adjustment, 28.4% remained in the original state, and 13.8% transferred to "key intervention", indicating that the system has a strong ability to correct the deviation of the intermediate state samples. From the perspective of the overall distribution change, the proportion of "stable completion" in the test set was 35.5% before optimization, and increased to 62.7% after optimization. The proportion of "key intervention" decreased from 26.8% to 11.2%. The results show that the rhythm segmentation tips, action correction suggestions and exercise sequence adjustment output by the optimization module can effectively promote the transfer of teaching status to a better interval.

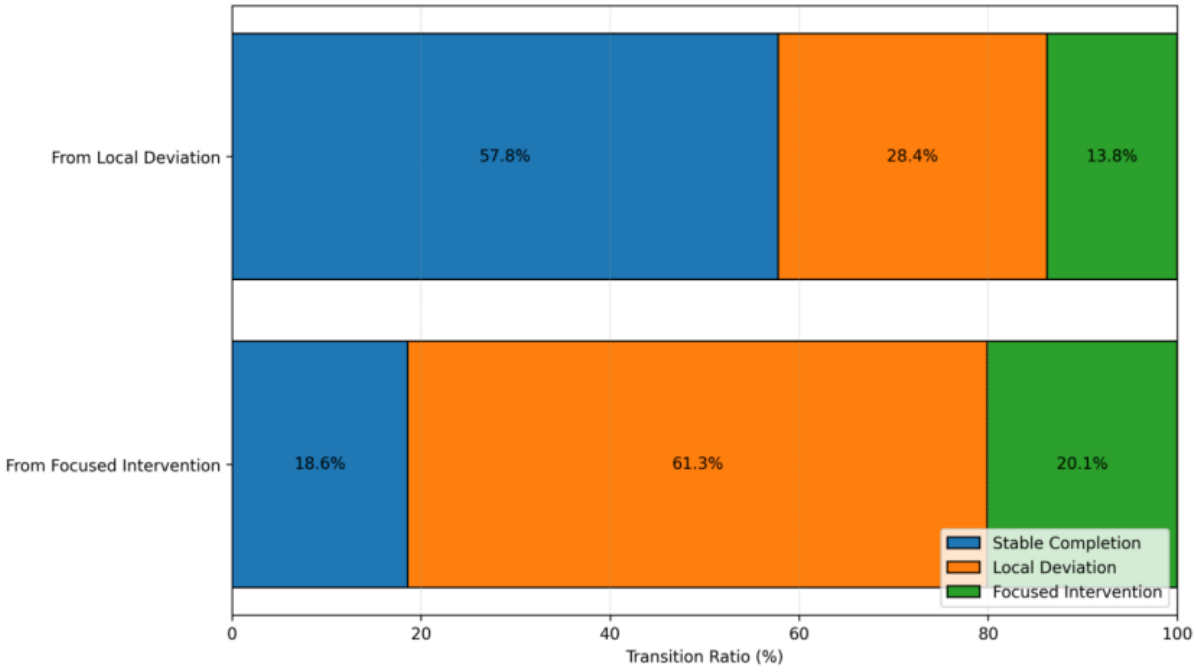

Figure 5: Scale diagram of state transition in traditional dulcute teaching

The results show that the exercise sequence rearrangement, rhythm segmentation tips and action correction suggestions output by the optimization module have a clear role, which can promote the convergence of teaching status to a better interval within a limited number of rounds, provide reliable basis for subsequent visual feedback and application adaptation, and maintain good consistency of result review.

4.4 Comparative experiment and visual result display

In order to further verify the comprehensive performance of the proposed method in the traditional dulcute teaching scene, the constructed model is compared with the Temporal MLP, BiGRU-Align, Single-Modal-CNN and dual-branch non-optimization models under the same data division and training conditions. The specific contribution of each module to the results is analyzed in combination with ablation experiments. The comparison metrics include Accuracy, Macro-F1, average inference delay, and state consensus rate. The results show that the model in this paper maintains a good level on the four indicators, especially after the linkage of multi-modal fusion and optimization generation, the fine-grained differences in the teaching fragments can be more completely retained. At the same time, although the two-branch non-

optimization model has the basic discrimination ability, the state boundary of the complex class shows greater fluctuations, and the difference is more obvious.

As shown in Fig. 6, box plots are used in this paper to show the Macro-F1 distribution of different models in five repeated rounds of experiments. The median of the model in this paper reaches 0.918, and the upper and lower interquartile ranges are 0.910 and 0.926, respectively, which are significantly smaller than the 0.861 to 0.889 of Temporal-MLP and 0.874 to 0.901 of BiGRU-Align. The median of Single-Modal-CNN is only 0.846. This result shows that the multimodal input and joint training mechanism not only improve the recognition accuracy, but also reduce the output dispersion between different batches.

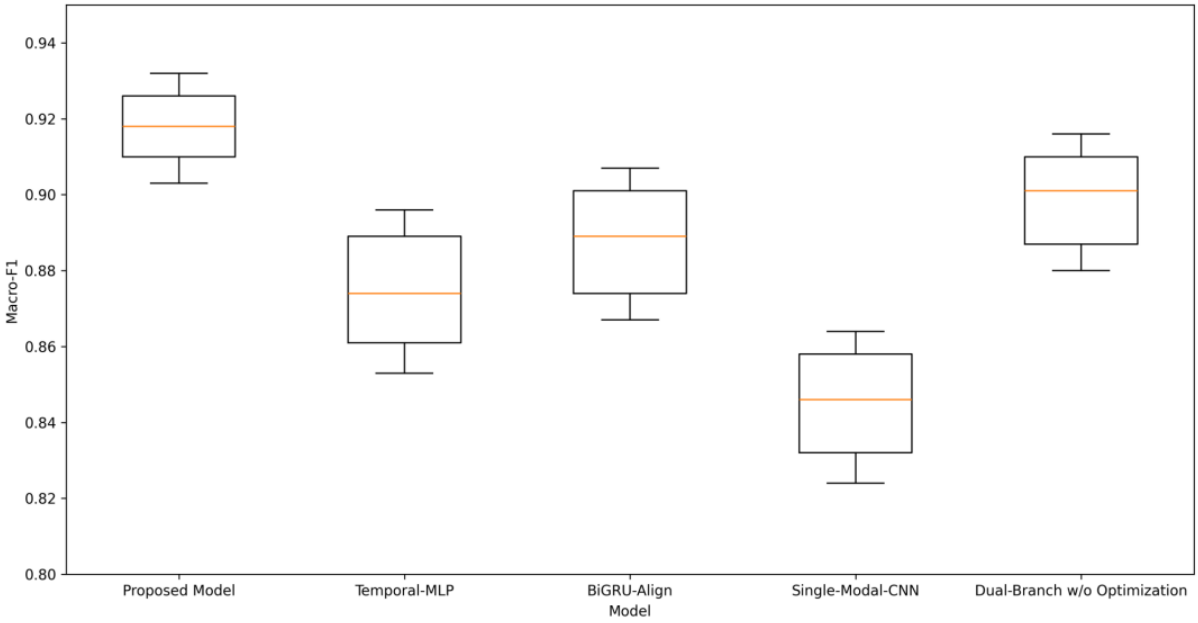

Figure 6: Box plots of Macro-F1 distribution for different models

In the module contribution analysis, this paper further removes key components for ablation testing. The complete model achieves 93.6% Accuracy, 0.918 Macro-F1, 24.6 ms average inference delay and 0.907 state consistency rate. After removing the action branch, the Accuracy decreased to 90.8%, and Macro-F1 decreased to 0.884. After removing the classroom behavior branch, the state consistency rate dropped to 0.861. After removing the optimization generation module, the inference delay is reduced to 22.1 ms, but Macro-F1 falls back to 0.892. The relevant results are shown in Table 3.

Table 3: Results of ablation experiments

Model Configuration	Accuracy / %	Macro-F1	Inference Latency / ms	State Consistency Rate
Full Model	93.6	0.918	24.6	0.907
Without Action Branch	90.8	0.884	23.4	0.873
Without Classroom Behavior Branch	91.5	0.891	23.7	0.861
Without Optimization Generation Module	91.9	0.892	22.1	0.879

In general, the proposed model achieves a good balance between recognition accuracy, state

cooperation and operation stability. The comparative experiments show that the multi-modal fusion structure enhances the ability to describe the differences in teaching performance. The ablation experiment further verifies the collaborative value of the action branch, the classroom behavior branch and the optimization generation module in the overall framework, which forms a complete support for the experimental analysis of the whole paper.

5 Conclusion and Prospect

5.1 Summary of Research

Focusing on the effect evaluation, system identification and optimization generation in traditional dulcute teaching, this paper constructs a complete computing link composed of data modeling, feature extraction, model training, result output and application adaptation. In the data layer, the synchronous collection and standardization organization of teaching audio, performance action and classroom process are completed, and a multi-dimensional label system and database structure for assessment tasks are established, so that information from different sources can enter a unified tensor space. In the model layer, this paper constructs a multi-modal teaching effect evaluation framework, which incorporates acoustic representation, action trajectory and classroom behavior coding into the same discrimination process, and enhances training stability through joint loss, dynamic scheduling and structure optimization. In the output layer, continuous scoring, state recognition and optimization suggestions are linked to realize structured feedback for teaching scenarios. The experimental results show that the constructed system achieves a balance between accuracy, state consistency and reasoning efficiency, can more stably identify stage differences and deviation characteristics in traditional dulctic teaching, and provide reliable support for classroom adjustment, training tracking and teaching adaptation. From the overall structure, this paper does not stop at the accuracy improvement level of a single model, but connects the teaching data organization, evaluation mechanism design, system state discrimination and result generation interface into a reusable information system framework. The framework not only retains the professional attributes of timbral details, chord percussive rules and classroom advancement rhythm in dulctic teaching, but also reflects the engineering value of artificial intelligence methods in the digital analysis of traditional instrumental music teaching, which lays a complete technical foundation for subsequent system expansion and platform deployment.

5.2 Limitations and Future research Directions

Although the current sample covers three levels of beginner, advanced and improved, the data sources are still mainly classroom training clips, and the coverage of cross-regional teacher styles, different instrument configurations and complex environmental recording conditions is not sufficient. Therefore, there is still room for continued expansion of the model's transfer ability in a wider range of teaching scenarios. At present, the action branch mainly relies on the trajectory of key points and local displacement statistics, and the description of the details of the harp and bamboo chord, the wrist micro-vibration and the short-term changes in the high-speed segments is still coarse-grained. Although the classroom behavior branch can reflect the process information such as demonstration, correction and repeated practice, it still lacks a deeper temporal modeling of the cumulative changes in the long-term learning path. The follow-up research can be promoted in three directions. First, the graph neural network and contrastive learning mechanism are introduced to enhance the ability of cross-segment structure expression and few-shot adaptation. Secondly, the multi-terminal database interface is extended to form a unified index structure of audio, video, log and comment to improve the efficiency of

system retrieval and playback. The third is to strengthen the lightweight deployment and online update design, so that the model can maintain continuous operation and stable feedback in the teaching platform, mobile terminal and auxiliary exercise scene. In addition, the subsequent system can also combine federated learning and privacy computing strategies to complete cross-school collaborative training without directly aggregating the original teaching data, and improve the transparency of the use of the teacher end and the learning end through the adaptive evaluation threshold and interpretable visualization module. The intelligent evaluation, state recognition and optimization generation in traditional dulcimer teaching form a more stable closed-loop structure landing application.

References

[1] Cui K. Artificial intelligence and creativity: piano teaching with augmented reality applications[J]. Interactive Learning Environments, 2023, 31(10): 7017-7028.

[2] Chen W. Research on the design of intelligent music teaching system based on virtual reality technology[J]. Computational intelligence and neuroscience, 2022, 2022(1): 7832306.

[3] Yuan K. Research on music teaching systems assisted by artificial intelligence[J]. International Journal of e-Collaboration (IJeC), 2024, 20(1): 1-17.

[4] Chen Y, Sun Y. The usage of artificial intelligence technology in music education system under deep learning[J]. Ieee Access, 2024, 12: 130546-130556.

[5] Song L. Design and implementation of remote piano teaching based on attention-Induced Multi-head Convolutional neural network optimized with Hunter–Prey optimization[J]. International Journal of Computational Intelligence Systems, 2024, 17(1): 2.

[6] Wang X. Design of vocal music teaching system platform for music majors based on artificial intelligence[J]. Wireless Communications and Mobile Computing, 2022, 2022(1): 5503834.

[7] Chen S. The application of big data and fuzzy decision support systems in the innovation of personalized music teaching in universities[J]. International Journal of Computational Intelligence Systems, 2024, 17(1): 215.

[8] Wang X. Deep learning-based personalized learning recommendation system design for" T++" Guzheng Pedagogy[J]. International Journal of Information Technology, 2024, 16(5): 2775-2781.

[9] Su Y, Wang Y. Optimization of music education strategy guided by the temporal-difference reinforcement learning algorithm: Y. Su, Y. Wang[J]. Soft Computing, 2024, 28(13): 8279-8291.

[10] Fang Q. Research on VR e-learning based on load balancing algorithm in intelligent music education system[J]. Entertainment Computing, 2024, 50: 100712.

[11] Jiang L. A fuzzy clustering approach for cloud-based personalized distance music education and resource management[J]. Soft Computing, 2024, 28(2): 1707-1724.

[12] Chen Y. Interactive piano training using augmented reality and the Internet of Things[J]. Education and Information Technologies, 2023, 28(6): 6373-6389.

[13] Huang N, Ding X. Piano music teaching under the background of artificial intelligence[J]. Wireless Communications and Mobile Computing, 2022, 2022(1): 5816453.

[14] Ji C, Wang D, Wang H. An investigation on the application of deep reinforcement learning in piano playing technique training [J]. Applied Mathematics and Nonlinear Sciences, 2024, 9(1).

[15] Ghatas Y, Fayek M, Hadhoud M. A hybrid deep learning approach for musical difficulty estimation of piano symbolic music[J]. Alexandria Engineering Journal, 2022, 61(12): 10183-10196.

[16] Ramoneda P, Jeong D, Eremenko V, et al. Combining piano performance dimensions for score difficulty classification[J]. Expert Systems with applications, 2024, 238: 121776.

[17] Enkhbat A, Shih T K, Cheewaprakobkit P. Human action recognition and note recognition: a deep learning approach using STA-GCN[J]. Sensors, 2024, 24(8): 2519.

[18] Koh E Y, Cheuk K W, Heung K Y, et al. MERP: A music dataset with emotion ratings and raters' profile information[J]. Sensors, 2022, 23(1): 382.

[19] Medina Y O, Beltrán J R, Baldassarri S. Emotional classification of music using neural networks with the MediaEval dataset[J]. Personal and Ubiquitous Computing, 2022, 26(4): 1237-1249.

[20] Louro P L, Redinho H, Malheiro R, et al. A comparison study of deep learning methodologies for music emotion recognition[J]. Sensors, 2024, 24(7): 2201.

[21] Jiang X, Zhang Y, Lin G, et al. Music emotion recognition based on deep learning: A review[J]. IEEE Access, 2024, 12: 157716-157745.