ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Research on Teaching Quality Evaluation of Music Course empowered by artificial Intelligence

Zijian Wang[1,*] *and Xiaodong Fu*[1]

[1] Department of Musicology, China Conservatory of Music, Beijing 100875, China

SUMMARY: *The integrated development of artificial intelligence and music education has put forward new requirements for objective and computable teaching quality evaluation. In this paper, we propose MusicEval-Net, a multimodal evaluation framework that fuses classroom audio, learner behavior logs, and textual feedback for teaching quality analysis of music courses. The system extracts pitch stability, rhythm deviation, sound intensity, interaction frequency, task completion, and semantic emotion features, which are mapped to teaching quality indicators through a multimodal deep learning model. Experiments were conducted in a university music course involving 120 students, 24 teaching sessions, and 1860 labeled samples. Compared with the manual scoring and the unimodal baseline, the Accuracy of the proposed model reaches 91.8%, the F1-score is 0.903, the MAE is 0.217, and the Cohen's κ is 0.84. The model maintains stable results in terms of performance evaluation, participation recognition and feedback consistency judgment, and is incorporated into the manual review link, providing a computational path for scalable, data-driven and traceable teaching quality evaluation of music courses.*

KEYWORDS: *Artificial intelligence; Music course; Teaching quality evaluation; Multi-modal feature fusion*

1 Introduction

Artificial intelligence pushes music curriculum evaluation from experience judgment to computable data modeling. The singing intonation, rhythm stability, strength control, work understanding, classroom interaction and practice completion in the music class not only contain audio signals, but also contain learning behavior logs and text feedback. Traditional evaluation relies on teachers' on-site listening and after-class records, which can reflect artistic judgment, but it is difficult to maintain a consistent scale in multi-class, multi-class and multi-task situations. With the development of deep learning and multimodal learning analysis, teaching quality evaluation can form evidence through audio recognition, behavior sequence modeling and semantic analysis, so that classroom performance, teaching feedback and learning outcomes can enter the same computational framework.

Related research provides a methodological basis for the intelligence of music evaluation. Kokturk-Guzel et al. studied the automatic assessment of students' rhythmic imitation performance and proved that rhythm deviation, timing matching and imitation stability could be quantified by digital signal processing methods [1]. Faghih et al. proposed a start-end point detection method for real-time and offline scenes of singing, so that the classroom singing clips could be accurately segmented and provide boundaries for the analysis of pitch,

*18335163369@163.com
https://doi.org/10.65102/is20261046 Published: 10 June, 2026

ventilation and rhythm [2]. Systems such as Civit have sorted out the scope and applications of AI music generation, indicating that deep models have the ability to process symbolic music, audio structure and style features [3]. Ferreira et al. investigated the generation of symbolic music using deep learning models and showed that melodic, rhythmic, and harmonic structures can be transformed into trainable sequence features [4]. Paroiu and Trausan-Matu proposed to use deep neural networks and dissonance to measure music aesthetics, which provided aesthetic feature modeling ideas for computational assessment of music performance quality [5]. Louro et al. compared the performance of various deep learning methods in music emotion recognition, and proved that convolutional network, recurrent network and Transformer structure have different advantages in emotional feature extraction [6]. Modran et al. used deep learning to identify musical emotions and therapeutic effects, indicating that a learnable mapping could be established between sound features and emotional responses [7]. Kang et al. proposed a video soundtrack generation method based on emotional multimodal Transformer, demonstrating the modeling value of visual, emotional and musical feature fusion [8]. Kwiecień et al. analyzed the technical, musical and legal attributes of artificial intelligence-assisted algorithmic music production systems, which provided reference for the deployment and compliance of musical intelligent systems [9].

Existing works focus on rhythm imitation recognition, singing segment detection, music generation or emotion recognition, and the evaluation objects are mainly sound works or algorithm outputs. The teaching quality evaluation of music courses involves teachers 'teaching organization, students 'classroom participation, work performance process, practice feedback and evaluation consistency, and the data sources are more complex. The single audio model can only describe the sound performance, and it is difficult to explain the classroom interaction and learning feedback. Single questionnaire or performance analysis lack of reviewable process evidence. Without multi-modal fusion, index mapping and model checking mechanism, intelligent evaluation is easy to stay at the result statistics level, which is difficult to meet the requirements of technical journals for algorithm structure, experimental verification and system implementation.

Based on this, this paper focuses on the artificial intelligence empowered music course teaching quality evaluation, and constructs an intelligent evaluation model that integrates audio, behavior and text features. The model takes class recordings, student practice logs, teacher scoring records and learning feedback texts as input, extracts features such as pitch deviation, rhythm stability, volume change, interaction frequency, task completion rate and semantic emotion, and completes the mapping of teaching quality indicators through the deep learning network. The system connects to the course platform at the interface layer, records the evaluation process, model output and manual verification results, introduces interpretability analysis and fairness detection, and reduces the evaluation deviation between voice part, basic level and classroom tasks. The evaluation framework formed in this paper can transform the sound performance, learning behavior and text feedback in music classroom into computable features, and make the teaching quality evaluation shift from a single manual judgment to a model analysis supported by multi-source data.

2 Related work

2.1 Application of Artificial Intelligence in music Education evaluation

After artificial intelligence enters the music education evaluation scene, the course quality no longer only depends on teachers 'subjective listening and final score records, but gradually turns to the computational evaluation supported by audio signals, performance behavior and

classroom feedback. Music learning is characterized by the coexistence of continuous performance, instant correction and emotional expression. The pitch, rhythm, strength, ventilation, interaction times and practice trajectories generated in the class can be transformed into model input.

Shahriar studied the application of generative adversarial networks in text generation of visual art, music and literature, and showed that the generative model can learn style patterns and structure distribution from high-dimensional art data, which provides an algorithm reference for work style recognition and performance quality modeling in music course evaluation [10]. Martinez-Roig et al. studied the application of social robots in music education and pointed out that robots can participate in beat prompt, action guidance and instant feedback, which provides a scene basis for classroom evaluation system to collect interactive behavior, learning response and performance process data [11]. Kasneci et al. studied the opportunities and risks of large language models in education and showed that language models can assist in interpreting learning content, generating feedback texts and analyzing learning logs, but the evaluation results still need to be combined with course objectives and manual review [12]. Akgun and Greenhow studied the ethical challenges of educational application of AI in K-12 scenarios, emphasizing that data privacy, algorithm bias, and evaluation transparency should be embedded in the system design process [13].

The above studies collectively show that artificial intelligence has been able to process sound, behavior and text information in music courses, but it is still necessary to transform individual technical capabilities into stable index mapping links for teaching quality evaluation. The music classroom evaluation system should extract the pitch and rhythm deviation in the audio recognition layer, record the practice persistence and interaction density in the behavior analysis layer, identify the feedback tendency and cognitive state in the text semantic layer, and then complete the quality score by the deep model.

Different from ordinary performance evaluation, music curriculum quality evaluation pays more attention to classroom process evidence and performance change trajectories. Therefore, the model needs to retain the professional boundaries of artistic judgment, while providing reviewable feature sources and scoring basis. In the actual system, classroom recordings can be sliced by bars, phrases and task nodes, student behavior logs can be aggregated by time Windows, and teacher comments can be transformed into feedback vectors by semantic encoding. When the multi-source features enter the unified evaluation model, they can reduce the fluctuation of single listening score, so that the same work can maintain a closer evaluation scale between different classes, different classes and different teachers. This kind of method does not weaken teachers 'judgment, but precipitate teachers 'experience into trainable labels, so that music curriculum quality evaluation has the basis of data tracking, process playback and model iteration. This evaluation path is more suitable for large-scale course platforms, and it is also convenient for subsequent comparative experiments and continuous verification across grades and types of works.

2.2 Multimodal learning analysis and intelligent evaluation of teaching quality

Multimodal learning analysis provides a finer technical path than single score statistics for intelligent evaluation of teaching quality. The learning state in the music classroom is reflected not only in the sound performance, but also in the eye movement attention, action response, platform click, assignment submission and teacher feedback text. If only the final test or questionnaire results are used, the evaluation model is difficult to capture the dynamic changes of students in rehearsal, model singing, rhythm training and work analysis.

Hlosta et al. studied predictive learning analytics in online education and showed that

model evaluation cannot only pursue score fitting, but also need to identify error sources and feature contributions [14] by explaining the algorithm misunderstands the learning outcome prediction process [14]. Lamb et al. studied the real-time prediction of learning results based on hemodynamic signals in virtual reality and online learning, and proposed the idea of using machine learning to classify learning states, which provided reference for the linkage analysis between physiological reactions, attention states and performance results in music classroom [15]. Laupichler et al. studied the artificial intelligence literacy in higher education and adult education, and pointed out that learners' understanding of intelligent systems would affect technology acceptance and learning behavior. Therefore, the intelligent evaluation of music courses should incorporate system readability and feedback understandability into the design [16]. Caspari-Sadeghi studied machine learning methods in technology-enhanced evaluation, combing the application of classification, regression, automatic scoring and feedback generation in evaluation scenarios, indicating that intelligent evaluation should pay attention to model accuracy, interpretability and usage boundaries at the same time [17].

Based on these studies, three input channels of audio, behavior and text can be constructed for music teaching quality evaluation. The audio channel is used to identify pitch fluctuation, rhythm offset, duration stability and timbre variation. The behavioral channel recorded the number of exercises, interaction frequency, classroom response and task completion time. The text channel analyzes the semantic tendency in student self-evaluation, teacher comment and system feedback. The multi-modal fusion layer can use the attention mechanism to calculate the weights of various features in different evaluation tasks to avoid excessive influence of a single modality on the scoring results.

For different course tasks such as vocal music, instrumental music and comprehensive music appreciation, the model can also set differentiated index mapping to make singing quality, practice engagement, work understanding and classroom participation into a unified scoring space. This evaluation method can generate traceable evidence of teaching quality, retain the expressive characteristics of music classroom, and also meet the requirements of intelligent systems for structured data, algorithm verification and scalable deployment.

In the model implementation, different modalities can be temporally aligned first, and then enter the shared representation space through feature projection. The class audio retained singing details with frame-level features, the behavior log recorded learning rhythms with event sequences, and the text feedback presented cognitive tendencies with semantic embedding. The fusion results are then entered into regression or classification heads to output teaching quality scores, participation levels, and feedback consistency indicators, and to facilitate experimental review.

2.3 Deficiencies of existing studies

The existing research has provided the algorithm, data and system basis for intelligent evaluation, but there is still a distance between it and the teaching quality evaluation of music courses. Martinez-Comesana et al. studied the influence of artificial intelligence on the evaluation methods of primary and secondary schools, and pointed out that automatic scoring, learning diagnosis and process feedback are changing the traditional evaluation structure [18]. Ouhaichi et al. studied the development trend of multimodal learning analysis and emphasized that video, audio, sensors and platform logs can jointly describe the learning process [19]. Ashwin et al. studied an intelligent tutoring system based on computer vision to detect body movements, indicating that action recognition can be used to judge learners' participation status and task completion [20]. Yildirim-Erbasli and Bulut proposed a conversational evaluation method, which used digital formative evaluation to stimulate test

engagement, and provided ideas for the combination of text interaction and evaluation feedback [21]. The relevant results are valuable for this paper, but it still needs to be organized with the sound performance, classroom tasks and teaching evaluation indicators of music courses.

In order to further clarify the corresponding relationship between relevant research and teaching quality evaluation of music courses, this paper is sorted out from four aspects: research methods, application scenarios, referable content and main limitations. The specific comparison results are shown in Table 1.

Table 1: Comparison of related studies on evaluation of AI education

References	Research method	Application scenario	Useful reference	Main limitation
Martinez-Comesana et al. [18]	Systematic review and summary of assessment methods	Intelligent assessment in primary and secondary education	Summarizes the influence of artificial intelligence on assessment methods	Insufficient integration with musical performance data
Ouhaichi et al. [19]	Systematic mapping of multimodal learning analytics	Learning process data modeling	Emphasizes the integration of audio, video, behavior, and other data	Lacks indicator mapping for music courses
Ashwin et al. [20]	Review of computer vision-based body movement recognition	Intelligent tutoring systems	Provides ideas for body movement detection and learning support	Weak in analyzing sound performance
Yildirim-Erbasli and Bulut [21]	Conversation-based formative assessment	Digital assessment scenarios	Focuses on interactive feedback and test-taking effort	Difficult to cover singing and instrumental performance processes

Table 1 shows that related studies cover automatic evaluation, multimodal learning analysis, visual action detection and conversational evaluation, but most of them are oriented to general learning tasks and lack joint modeling of acoustic performance, performance behavior and aesthetic feedback in music classroom. The teaching quality evaluation of music courses should not only be based on grades or participation times, but also need to analyze the stability of intonation, accurate rhythm, timeliness control, work understanding and classroom interaction in students 'singing.

Existing intelligent evaluation methods tend to favor platform behavior or text question answering in feature selection, and pay insufficient attention to the consistency of audio segment segmentation, acoustic feature extraction and music task labels. Although some studies use multi-modal data, they often stay in parallel input between different modalities, and lack weight allocation and semantic alignment for teaching quality indicators. For music courses, the same singing may present different evaluation results in intonation, rhythm, expressiveness and classroom task completion. Without an interpretable mechanism in the model, it is difficult for teachers to judge whether the score is from sound quality, behavioral input or text feedback.

The existing research rarely discusses the influence of voice type, basic level, track difficulty and device acquisition differences on model output in terms of fairness. Based on these shortcomings, this paper integrates audio recognition, behavior log analysis and text

semantic modeling into a unified framework, constructs an interpretable teaching quality scoring model, and adds manual verification, deviation detection and result tracking mechanisms in the system deployment, so that the evaluation results have reviewable data sources and stable application boundaries.

3 Methods

3.1 The construction of teaching quality evaluation model based on the fusion of audio, behavior and text features

This paper constructs a multimodal intelligent model for music teaching quality evaluation. The model takes class recordings, student platform behaviors, teacher comments and student feedback texts as input, and maps pitch, rhythm, intensity, classroom interaction, task completion and semantic orientation into a unified scoring space. The overall structure does not adopt a single performance prediction, but converts teaching organization, learning participation and music performance into computable evidence, so that the evaluation results have data sources, model paths and manual verification records.

Fig. 1 shows the overall operational link of the model. The audio end collects classroom singing, rhythm imitation and instrumental music practice clips, the behavior end records login, click, practice times, submission time and interaction frequency, and the text end receives teacher comments, student self-evaluation and system feedback. After entering the preprocessing module, the three types of data complete denoising, slicing, time alignment and anonymization, and then are sent to the acoustic encoder, behavior sequence encoder and text semantic encoder respectively. The fusion layer outputs the teaching quality score, participation level and feedback consistency results, and the verification layer saves the manual review marks for subsequent training updates.

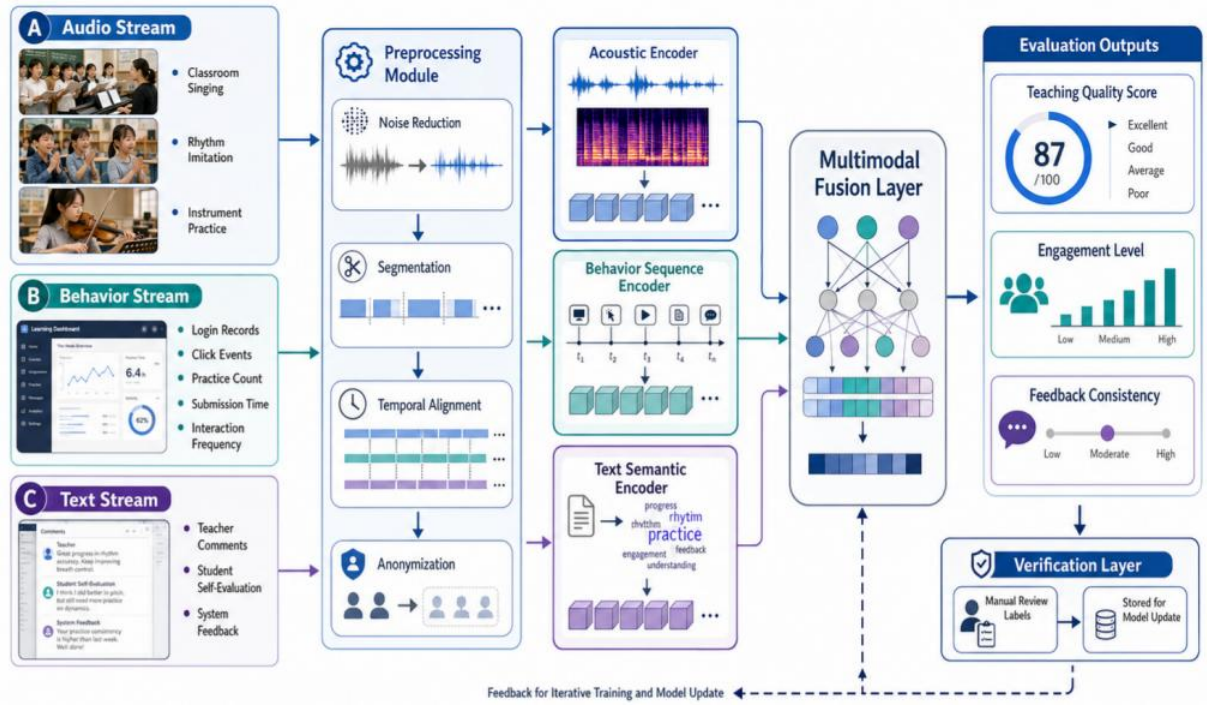

Figure 1: Framework of multimodal evaluation model for teaching quality of music courses

In order to unify the input scale of classroom audio behavior log and text feedback, three kinds of feature coding and splicing preprocessing processes should be completed first, and the specific process is shown in the following equation:

$$u_i = \sigma(\Theta_f[\phi_a(A_i) \oplus \phi_b(B_i) \oplus \phi_t(T_i)] + \beta_f) \tag{1}$$

Here,A_i represents the audio sample of the i classroom segment, B_i represents the behavior log in the same time window, and T_i represents the text feedback. $\phi_a(\cdot)$, $\phi_b(\cdot)$, $\phi_t(\cdot)$ denote acoustic, behavioral, and textual encoding functions, respectively. $\oplus$ denotes the vector concatenation, Θ_f and β_f are the fused projection parameters, σ is the nonlinear activation function, and u_i is the normalized multimodal base representation. This formula is used to solve the input scale difference caused by the inconsistency of three types of data sampling frequency, structure granularity and semantic density.

In order to avoid that a single modality occupies too high weight in the scoring, the system introduces the cross-modal attention coefficient to calculate the contribution weight of each feature, and the calculation process is shown as follows:

$$\alpha_m = \frac{\exp\left(q^\top \tanh(\Theta_m u_{i,m} + \beta_m)\right)}{\sum_{r\in\{a,b,t\}} \exp\left(q^\top \tanh(\Theta_r u_{i,r} + \beta_r)\right)}, \quad z_i = \sum\nolimits_{m\in\{a,b,t\}} \alpha_m u_{i,m} \tag{2}$$

Here, m represents the audio, behavior or text modality, α_m represents the attention weight of the corresponding modality in the current evaluation task, q is the task query vector, Θ_m and β_m are used to complete the modal transformation, and z_i is the fused evaluation representation. This formula enables the model to automatically adjust the feature contribution under different tasks such as rhythm training, singing performance and classroom participation, avoiding the scoring results being dominated by a certain type of data.

In order to map the fusion representation to the teaching quality score and constrain the consistency relationship between the model output and the manual annotation, the joint loss function is established as follows:

$$\hat{y}_i = \rho(\Gamma_2 \delta(\Gamma_1 z_i + \eta_1) + \eta_2), \quad \mathcal{L}_q = \frac{1}{n}\sum\nolimits_{i=1}^{n} |y_i - \hat{y}_i| + \lambda(1-\kappa) \tag{3}$$

Here, $\hat{y}_i$ represents the teaching quality score predicted by the model, y_i represents the teacher labeling result, Γ_1, Γ_2, η_1,η_2 are the score mapping parameters, ρ is used to limit the output range, δ is the hidden layer activation function, κ represents the consistency coefficient between the model score and the manual score, and λ is the consistency constraint weight. The formula simultaneously constrains the scoring error and manual consistency, so that the model output not only has numerical accuracy, but also retains the professional judgment boundary in music curriculum evaluation.

In the training sample construction, the classroom segments are doubly labeled according to the teaching task and the work structure, the phrase boundary is reserved for the vocal task, the beat point position is reserved for the rhythm task, and the discussion round is reserved for the appreciation task. Double-teacher review was used for manual scoring, and the system synchronously recorded scoring differences, feature contributions, and model versions. After this processing, the evaluation result is no longer just a single score, but a link record composed of original data, feature coding, fusion weight and manual confirmation, which can support subsequent experimental comparison and classroom application tracking. For different students in the same classroom, the model generates individual representations

according to task Windows, and then summarizes them into class-level quality portraits, which facilitates the observation of the correspondence between teachers 'organizational rhythm and student performance changes. At the same time, it retains a complete, stable and backtrackable calculation basis.

3.2 Deep learning training and index mapping Algorithm for Music Course evaluation task

The training objective of the music course evaluation task, is to further feed the fused features obtained in the previous section into a deep network, so that the model can simultaneously identify the relationship between sound performance, learning engagement, and textual feedback. In this paper, the class clips are divided into training set, validation set and test set with a ratio of 7∶1.5∶1.5. The training samples included four types of tasks, including vocal singing, rhythm imitation, instrumental music practice and classroom discussion. Each sample was bound to the teacher's rating, task type, track difficulty and student's basic label to avoid the model only learning a single score result.

Fig. 2 illustrates the training and metric mapping process. After the samples enter the training end, the batches are organized according to the task type, and then the high-level representation is extracted by the deep encoding network. The loss calculation layer simultaneously bound the quality scoring error, participation level error, and feedback consistency error. The verification end completes the level mapping according to the teacher's rating scale, and writes the prediction score, confidence and deviation tags into the result cache for subsequent system deployment and call.

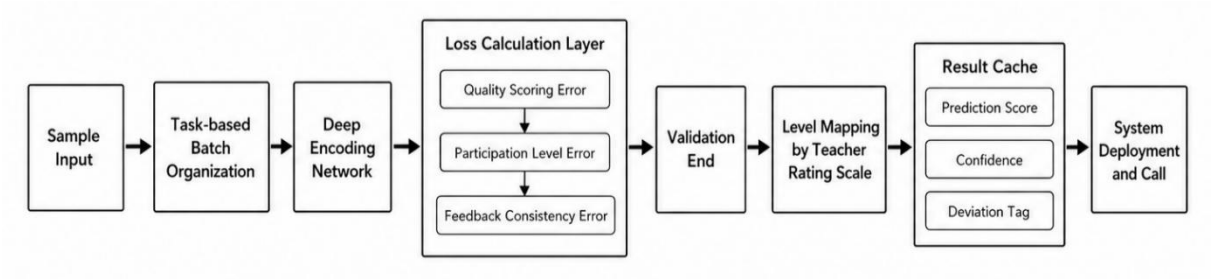

Figure 2: The training and indicator mapping process of the intelligent evaluation model for music courses

In order to make the intonation rhythm strength and the behavior sequence participate in the training together and keep the acoustic evidence stable, this section constructs a weighted multi-task loss function, as shown in the following equation:

$$\mathcal{L}_m = \omega_1 \frac{1}{N}\sum_{i=1}^{N}(\hat{s}_i - s_i)^2 + \omega_2 \sum_{c=1}^{C} - g_{i,c}\log \hat{g}_{i,c} + \omega_3 \frac{1}{N}\sum_{i=1}^{N}|r_i - \hat{r}_i| \qquad (4)$$

where N represents the number of training samples, s_i and $\hat{s}_i$ represent the teacher quality score and model prediction score, $g_{i,c}$ and $\hat{g}_{i,c}$ represent the true label and predicted probability of the ch participation level,r_i and $\hat{r}_i$ represent the feedback consistency label and predicted value, ω_1 to ω_3 are the task weights. This loss function puts continuous ratings, classification ratings, and feedback consistency within the same training objective, enabling the model to fit both music performance scores and identify classroom participation and feedback stability.

In order to transform the continuous predicted values into interpretable course evaluation

levels and corresponding to teacher rating scales, the system establishes a threshold indicator mapping function, as shown in the following equation:

$$Q_i = \begin{cases} A, \hat{s}_i \geq \tau_3, c_i \geq \mu \\ B, \tau_2 \leq \hat{s}_i < \tau_3 \\ C, \tau_1 \leq \hat{s}_i < \tau_2 \\ D, \hat{s}_i < \tau_1 \end{cases} \tag{5}$$

Among them, Q_i represents the teaching quality level of the i sample, A, B, C and D correspond to excellent, good, qualified and to be improved respectively, τ_1 to τ_3 are the scoring thresholds, c_i represents the model output confidence, μ represents the credible judgment threshold. The mapping function enables the model results to correspond with the teacher evaluation scale, and reserves the low-confidence samples for manual review, reducing the instability caused by automatic scoring directly entering the conclusion.

In order to reduce the score drift caused by the distribution difference of samples in different classes and constrain the output confidence interval, the confidence calibration constraint is added in the training phase, as shown in the following equation:

$$\mathcal{L}_{cal} = \mathcal{L}_m + \lambda_1 \sum_{k=1}^{K} |Acc(B_k) - Conf(B_k)| + \lambda_2 \|\Theta\|_2^2 \tag{6}$$

Here, B_k is the k interval divided by confidence, $Acc(B_k)$ is the true agreement rate of the prediction results in the interval, $Conf(B_k)$ is the average confidence, Θ is the set of model parameters, and λ_1 and λ_2 are the constraint weights. This formula is used to calibrate the model output probability to keep the high-scoring samples matched with the high-confidence outputs, and to suppress the rating fluctuations caused by the differences in the recording environment and curriculum tasks of different classes.

In the training process, AdamW is used as the optimizer, the initial learning rate is set to 0.0002, the batch size is 32, and the maximum number of training rounds is 80. The training is stopped when there is no improvement in the validation set for 8 consecutive rounds. The model output not only contains the total score, but also synchronously generates audio contribution, behavior contribution, and text contribution proportion. This setting enables the evaluation results to return to the specific feature source, which is convenient for teachers to check whether the model judgment is in line with the classroom reality, and also provides a stable basis for subsequent ablation experiments and system feedback verification. After indicator mapping, the system retains the random seeds, parameter versions and validation curves for each training, and writes the abnormal samples to the double-check queue. The results of teacher review re-enter the label library for the next round of mini-batch update, so as to maintain a continuous correspondence between model training and course evaluation, and avoid implicit deviation of scoring rules in different teaching stages. At the same time, the system records the proportion of samples in each level, which supports the subsequent fairness test and result review.

3.3 Intelligent evaluation system deployment, interface design and feedback verification mechanism

The deployment of the intelligent evaluation system focuses on the collection of the classroom end, the calculation of the platform end and the review of the teacher end. The system sets four kinds of services in the music course platform: audio upload, behavior log, text feedback and score writeback. The front-end is responsible for collecting classroom

singing clips, rhythm training records, instrumental exercise video indexes and student feedback texts, and the back-end completes feature calling, model reasoning, result caching and teacher review. The interface adopts unified token authentication and hierarchical authority control. The student end only receives personal feedback, the teacher end can view the class aggregation results, and the management end retains model versions, data fingerprints and abnormal call records.

In the system deployment phase, in order to ensure the stable storage of classroom audio, behavior and text data in the interface layer, unified data encapsulation rules should be established and back-end verification should be completed, as shown in the following equation:

$$\mathcal{D}_i = \mathrm{Pack}(a_i, b_i, t_i, l_i), \quad \chi_i = \mathrm{Hash}(\mathrm{id}_i \| \mathrm{time}_i \| \mathrm{task}_i \| v_s) \tag{7}$$

Here, a_i, b_i, and t_i represent audio, behavior, and text data, respectively; l_i represents teacher label or course task label;$\mathcal{D}_i$ is the packet encapsulated at the interface layer; χ_i is the link check code; v_s represents the system version. This formula binds the multi-source data and the course context into the same request object, so that the back-end can identify the sample source, task type and version status.

In the multi-class synchronous usage scenario, in order to control the delay fluctuation system setup request caused by concurrent uploading of different terminals, the scheduling scoring rule keeps response as shown in the following equation:

$$R_j = \arg\min_{q_j} \sum_{e=1}^{E} (\xi_1 d_e + \xi_2 c_e + \xi_3 m_e) - \xi_4 p_j \tag{8}$$

Here, q_j represents the j request queue, d_e represents the waiting delay, c_e represents the computational load, m_e represents the cache occupancy, p_j represents the classroom task priority, and ξ_1 to ξ_4 are the scheduling coefficients. This formula is used to maintain the stability of inference service when multiple terminals access simultaneously, and to avoid large return difference of real-time evaluation results due to congestion.

After the teacher reviews and enters the feedback link, in order to verify the consistency between the model output and the manual results, the feedback calculates the sample credibility and marks the reflux state, as shown in the following equation:

$$V_i = \sigma(\gamma_1 |\hat{s}_i - s_i^r| + \gamma_2 (1 - c_i) + \gamma_3 \Delta_i) \tag{9}$$

Here, s_i^r Represents the teacher review score, $\hat{s}_i$ represents the model output, c_i represents the model confidence, Δ_i represents the historical deviation of similar samples, and V_i represents the feedback check value. This formula is used to determine whether the sample needs to enter the review queue. The higher the value is, the more obvious the difference between the automatic output and the manual result is, and the system will retain the sample as the subsequent training material.

After the manual review of the sample is completed, in order to make the sample return to the training queue in time, the system adopts the incremental parameter update rule and retains the historical version record, as shown in the following equation:

$$\Theta_{r+1} = \Theta_r - \eta \nabla_\Theta (\mathcal{L}_{\mathrm{new}} + \beta \mathcal{L}_{\mathrm{old}} + \zeta \|\Theta - \Theta_r\|_2^2) \tag{10}$$

Here, Θ_r represents the current model parameters, Θ_{r+1} represents the updated parameters, $\mathcal{L}_{\mathrm{new}}$ represents the review sample loss, $\mathcal{L}_{\mathrm{old}}$ represents the historical sample

retention loss, and η, β, ζ are the update coefficients. This formula makes the feedback data participate in the model modification, and restricts the drastic changes of parameters, so that the previous evaluation rules will not be directly rewritten by a small number of new samples.

After the evaluation results enter the archiving stage, in order to ensure the whole process traceability system, a link fingerprint record is constructed for each sample and the model version number is bound, as shown in the following equation:

$$\Omega_i = \{\chi_i, \hat{s}_i, Q_i, c_i, Rev_i, \Theta_r, api_i, time_i\} \tag{11}$$

Here, Ω_i represents the archive fingerprint of the i evaluation record, Q_i represents the grade result, Rev_i represents the teacher review status, api_i represents the interface origin, and $time_i$ represents the writing time. This structure enables each score to trace back to data, model, interface and manual confirmation, which facilitates subsequent statistical analysis, anomaly detection and version comparison. After the system is deployed, all evaluation results do not directly cover teacher judgments, but form a chain of evidence that can be reviewed.

In actual operation, the system generates fine-grained records by class fragments, and then summarizes them into class quality portraits by students, works and teaching tasks. The teacher side can view the audio contribution, behavior contribution and text contribution ratio, and can also add correction labels to low confidence samples. The backend timing compares the interface response, score distribution and review differences. If there is a centralized deviation of a certain type of track or a certain device to collect samples, the system will pause the automatic writeback and only retain the candidate evaluation results. Such deployment puts model reasoning, human judgment and data governance in the same closed loop, so that music curriculum evaluation has manageable engineering boundaries. The system log synchronization retains the exception cause for subsequent review and model iteration analysis.

3.4 Model interpretability and fairness detection algorithm for evaluation bias control

The model interpretability and fairness detection algorithm for evaluation bias control is mainly used to check whether the scoring basis is clear and whether there are systematic errors between different groups. Music course evaluation is affected by differences in voice type, basic level, track difficulty, recording equipment and classroom tasks. If the model only outputs the total score, it is difficult for teachers to judge the source of the score. In this paper, an explanation layer and a fairness detection layer are set up outside the scoring model to decompose the contributions of three types of features: audio, behavior and text, and group verification is established according to student basis, course task and collection environment.

In order to explain the contribution differences of each modality in music scoring and locate the source of influence of acoustic, behavioral and textual features, the model uses contribution decomposition method to output interpretable evidence, as shown in the following equation:

$$E_{i,m} = \frac{|F(x_i) - F(x_i^{(-m)})|}{\sum_{r\in\{a,b,t\}} |F(x_i) - F(x_i^{(-r)})| + \varepsilon} \tag{12}$$

where $E_{i,m}$ represents the modal contribution of the m class in the i sample, $F(x_i)$ represents the scoring function under the full input, $x_i^{(-m)}$ represents the input after

removing a certain mode, and ε is used to prevent the denominator from being zero. This formula can show whether the rating is mainly from voice performance, behavior input or text feedback, so that the model output has a readable basis.

In order to measure the prediction error gap between different student groups and find the potential evaluation bias, the group fairness measurement index is constructed for the subsequent verification process, as shown in the following equation:

$$\Delta_{\text{fair}} = \max_{g \in G} |\,\text{MAE}_g - \overline{\text{MAE}}| + \lambda_d \max_{g \in G} |\,\bar{s}_g - \bar{y}_g| \tag{13}$$

Here, G represents the group set, MAEg repres artificial mean of the group, respMAE$_g$vely, λd is the mean difference constraigoefficie$\overline{\text{MAE}}$ This equation is used to compare the $\bar{s}_g$ error $\bar{y}_g$ evels of different groups and avoid the samples with weak base or poor recordin λ_d conditions being consistently underestimated.

In order to calibrate the relationship between the prediction confidence and the true agreement rate of each group, the system establishes the group confidence deviation constraint function, and retains the interval weight for verification, as shown in the following equation:

$$C_g = \sum_{k=1}^{K} \frac{|B_{g,k}|}{|B_g|} \left|\text{Acc}(B_{g,k}) - \text{Conf}(B_{g,k})\right| \tag{14}$$

Here, $B_{g,k}$ represents the k confidence interval in the g group, $\text{Acc}(B_{g,k})$ represents the true agreement rate in the interval, and $\text{Conf}(B_{g,k})$ represents the average confidence. This equation is used to check whether the model exhibits overconfidence on a certain class of students or tasks, so that the high confidence score can be kept close to the human review agreement rate.

In order to limit the bias propagation and maintain the scoring accuracy, this paper combines the base loss and fairness constraint as a joint training objective, and constrict the range of parameter drift, as shown in the following equation:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{base}} + \alpha_f \Delta_{\text{fair}} + \alpha_c \sum_{g \in G} C_g + \alpha_e \sum_{i=1}^{n} H\,(E_i) \tag{15}$$

Here, $\mathcal{L}_{\text{base}}$ represents the original scoring loss, α_f, α_c, α_e are the constraint weights, and $H(E_i)$ represents the contribution distribution entropy. This formula incorporates the fairness gap, confidence bias and explanation distribution into the training objective, so that the model retains a more balanced group performance in the pursuit of scoring accuracy.

In order to test the stable relationship between the interpretation results of the model and teachers 'professional judgment, the interpretation consistency similarity calculation is introduced and the comparison result value is reviewed, as shown in the following equation:

$$S_{\text{exp}} = \frac{\sum_m e_m^{\text{model}}\, e_m^{\text{teacher}}}{\sqrt{\sum_m (e_m^{\text{model}})^2}\sqrt{\sum_m (e_m^{\text{teacher}})^2}} \tag{16}$$

where e_m^{model} represents the proportion of modal contributions given by the model, e_m^{teacher} represents the proportion of basis recorded when the teacher reviews, and S_{exp} represents the interpretation consistency. This formula is used to judge whether the model explanation is

close to the teacher's professional judgment. If the similarity is low, the sample will enter the manual review cohort.

After detection, the system generates a bias report and an explanation summary, which includes group error, confidence interval shift, modal contribution proportion, and double-check sample size. Teachers can use this to determine whether the grading relies too much on audio quality or ignores classroom interaction and textual feedback. If the error of a group continuously exceeds the threshold, the system will reduce the weight of automatic scoring and increase the proportion of manual review. All interpretation results were written into the sample file synchronously, and the track, voice part, device and task labels were retained to provide a traceable basis for subsequent ablation experiments, error analysis and model correction.

4 Results and discussion

4.1 Experimental design of intelligent evaluation of music courses

This section focuses on the experimental environment, sample composition and evaluation process of MusicEval-Net. The experimental subjects are 120 students in two parallel classes of music courses in a university. The experiment period is 12 weeks, and 24 classroom records are collected, forming 1860 effective labeled samples. Sample sources include vocal singing, rhythm imitation, instrumental practice, and work discussion. Each sample saved the class audio, behavior log, teacher comments, and student self-evaluation text, and was independently scored by two teachers. The samples whose scores differed by more than 8 points entered the review process, and the final labeling consistency Cohen's κ was 0.84. In the data processing stage, the audio was resampled to 16kHz and segmented by phrase and task window. The number of exercises, interaction frequency, submission delay and completion rate were extracted from the behavior log. The feedback vector is obtained by semantic encoding for text data. The training set, validation set and test set were divided into 7∶1.5∶1.5, and stratified sampling was used to keep the task categories and rating levels consistent. The platform is implemented using Python3.10 and PyTorch2.1, the server is configured with RTX3090GPU, batch size 32, initial learning rate 0.0002, and maximum training rounds 80. The comparison models include Audio-CNN, Behavient-LSTM and Text-BERT, and the indicators are Accuracy, F1-score, MAE and consistency coefficient. The design incorporates model performance, evaluation stability and review results into the analysis link, which provides a unified basis for subsequent analysis and ablation.

4.2 Analysis of participation in music class and acceptance of intelligent evaluation

The experimental group used MusicEval-Net to submit classroom clips, view instant feedback and fill in self-assessment texts, while the control group continued to use the conventional platform to record and score teachers. The engagement data consisted of the completion rate of audio tasks, the number of uploading exercises, the frequency of classroom interaction, the feedback viewing rate and the self-evaluation submission rate. The technology acceptance consisted of the perceived usefulness, the perceived ease of use, the trust of evaluation and the clarity of feedback. All the behavioral indicators are from the system log, which does not rely on students to fill in the recall. The timestamp of the log is accurate to the second level, which can correspond to each singing, rhythm imitation and evaluation feedback.

To show the differences in participation between the two groups of students in the music

classroom task, Fig. 3 presents the five normalized indicators using a radar chart. In the experimental group, the scores of audio task completion rate, exercise upload times, classroom interaction frequency, feedback viewing rate and self-assessment submission rate were 0.92, 0.88, 0.81, 0.90 and 0.86, respectively. The control group were 0.79, 0.63, 0.58, 0.72 and 0.61, respectively. The difference in feedback viewing and practice uploading was more obvious in the experimental group, indicating that model feedback can link in-class singing, rhythm training and after-class practice. Compared with a single test, the continuous log can better reflect whether students re-practice according to the system prompts, and can also show the response process after the teacher's evaluation enters the student's behavior.

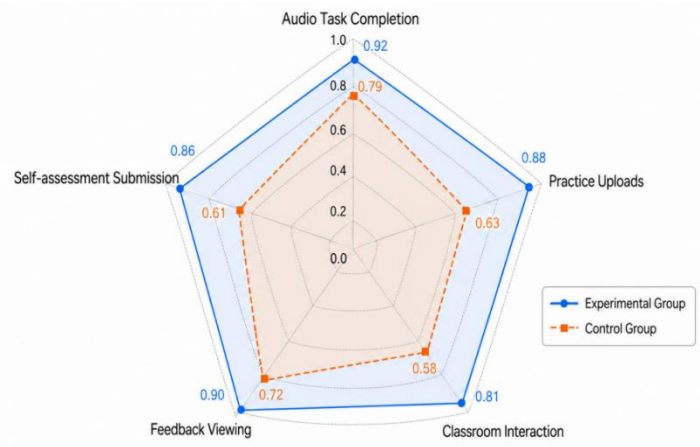

Figure 3: Radar chart of engagement in music class

To further observe the acceptance status of students to the intelligent evaluation results, Fig. 4 uses box plots to present the four questionnaire dimensions. In the experimental group, the median of perceived usefulness was 4.47, perceived ease of use was 4.31, evaluation trust was 4.18, and feedback clarity was 4.26. The corresponding values of the control group were 3.72, 3.80, 3.46, and 3.58. The interquartile ranges of the four dimensions are all less than 0.42, indicating that the acceptance distribution is relatively concentrated, and there is no phenomenon that a few high scores pull the overall results. Evaluation trust is lower than perceived usefulness, indicating that students recognize the value of model feedback, but it still needs the support of teacher review and result interpretation.

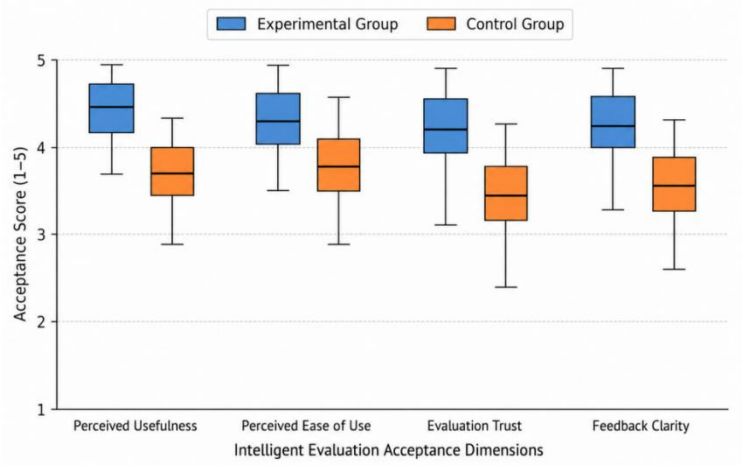

Figure 4: Boxplot of acceptance of intelligent evaluation

To keep the participation data corresponding to the platform behavior, Table 2 lists the main observation indicators of the experimental group and the control group. The results showed that the average number of login times per week in the experimental group was 5.8 times, which was significantly higher than 3.9 times in the control group. The feedback viewing rate reached 93.5%, 14.8 percentage points higher than that of the control group; The frequency of classroom interaction reached 4.1 times per class hour, indicating that students were more willing to adjust their singing and rhythm practice according to the system prompts.

Table 2: Comparison of music classroom engagement and acceptance of intelligent evaluation

Indicator	Experimental Group	Control Group	p-value
Weekly logins / times	5.8±1.1	3.9±1.3	<0.001
Audio task completion rate / %	92.4±4.8	80.6±6.5	<0.001
Feedback viewing rate / %	93.5±3.9	78.7±7.2	<0.001
Classroom interaction frequency / times	4.1±0.8	2.6±0.7	<0.001

Taking the above into account, it can be seen that MusicEval-Net has a strong traction effect on music classroom participation behavior. Students do not just complete the platform tasks, but form a continuous behavior chain between audio submission, feedback viewing, self-evaluation correction and secondary practice. The results show that the intelligent evaluation system can transform music performance data into specific feedback, and enhance students' acceptance of the evaluation results through visualization results and teacher review. The synchronous changes in engagement and acceptance provided a stable data basis for the subsequent analysis of the relationship between multimodal features and teaching quality ratings.

4.3 Correlation analysis and ablation experiment between multi-modal evaluation features and teaching quality scoring results

In this paper, the teaching quality score was split into three dimensions: vocal performance, learning engagement and feedback consistency, and the correlation coefficients between audio, behavioral and textual features and the total score were calculated. All features are from the test set output, and the training set samples are not used to avoid correlation being affected by the model fitting process. The correlation analysis was confirmed by Pearson coefficient and permutation test, the significance threshold was set at 0.05, and the feature values were standardized.

To show the strength of association between different features, Fig. 5 uses heat maps to present the correlation between core features and scoring results. The correlation coefficient between pitch stability and total score was 0.68, rhythm deviation was -0.64, task completion rate was 0.63, interaction frequency was 0.57, and semantic positivity was 0.51. Audio features had the strongest relationship with vocal performance ratings, behavioral features had a more stable relationship with learning engagement ratings, and text features mainly affected feedback consistency. The rhythm deviation was negatively correlated, indicating that the larger the error, the lower the score. The correlation coefficient of volume stability was 0.46, indicating that this index can provide complementary judgment, but it is not sufficient to determine the quality level alone.

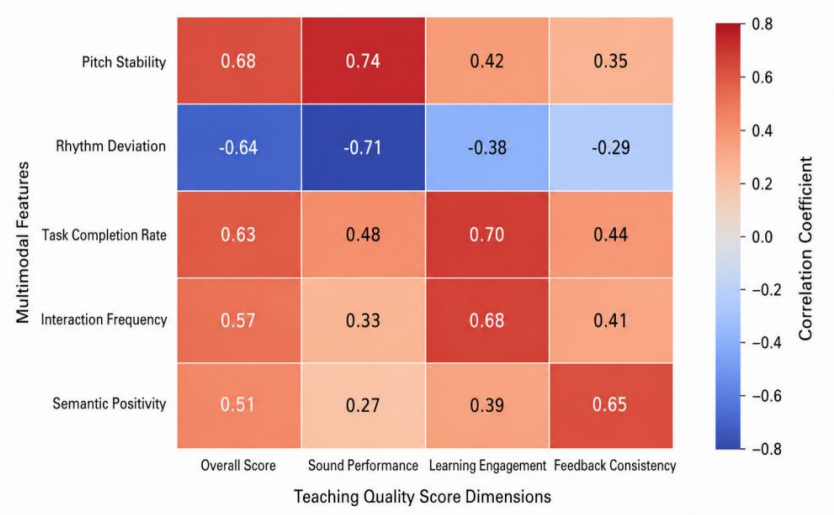

Figure 5: Heat map of correlation between multimodal features and teaching quality ratings

To compare modal contributions in different course tasks, Fig. 6 uses parallel coordinate plots to present the proportion of contributions in the four categories of vocal, rhythmic, instrumental, and appreciative tasks. In the vocal task, the audio contribution was 0.52, the behavior contribution was 0.26, and the text contribution was 0.22. For the appreciation task, the text contribution increased to 0.45 and the audio contribution decreased to 0.24. This result shows that the model does not rely on a certain kind of data, but adjusts the scoring basis according to the nature of the task. The behavior contribution reached 0.37 in instrumental music practice, which was closely related to multiple submission, beat error correction and segment retraining. The display behavior log could complement the process information that could not be covered by single audio quality.

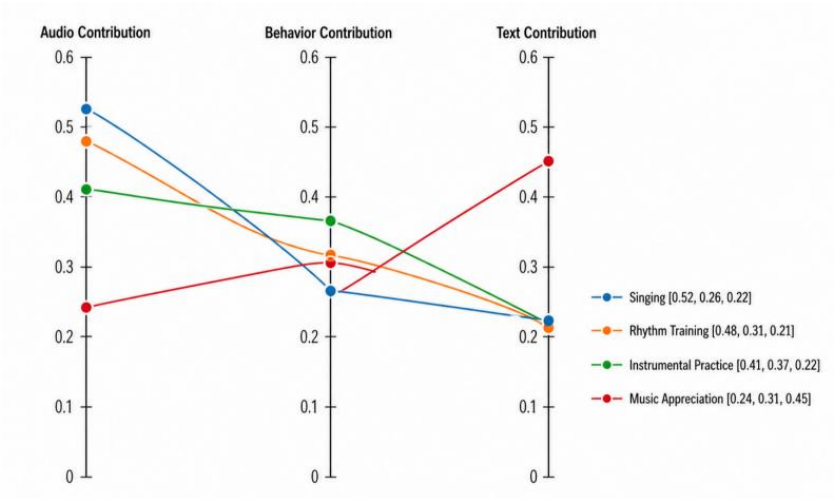

Figure 6: Parallel coordinate plots of modal contributions for different musical tasks

To verify whether the correlation results depend on a certain modality, ablation experiments are added to Table 3. The correlation coefficient and MAE of the full model were 0.74 and 0.217, respectively. After removing audio features, the correlation coefficient drops to 0.61, indicating that acoustic evidence is still the main support for music course evaluation. The MAE rose to 0.276 after removing the behavioral features, indicating that the learning process data had a supplementary effect on the stability of the score. After removing the text

features, the F1-score decreases to 0.861, and the feedback consistency recognition is affected.

Table 3: Results of ablation experiments for multi-modal evaluation features

Model Setting	Correlation Coefficient r	F1-score	MAE	Cohen’s κ
Complete MusicEval-Net	0.74	0.903	0.217	0.84
Without audio features	0.61	0.832	0.302	0.73
Without behavioral features	0.66	0.874	0.276	0.77
Without text features	0.69	0.861	0.251	0.79
Without attention fusion	0.64	0.846	0.289	0.75

Comprehensive experiments show that the teaching quality score is not determined by a single sound indicator, but is formed by audio performance, learning process and text feedback. Audio features provide the basis for singing and rhythm judgment, behavioral features reflect practice persistence and classroom response, and text features supplement student understanding and teacher feedback. The attention fusion mechanism can adjust the modal contribution according to the task type, so that the vocal, rhythmic, instrumental and appreciative tasks can obtain differentiated evaluation basis. The results provide a feature-level explanation basis for the subsequent model performance comparison, and also show that multimodal fusion is a necessary calculation link to maintain the scoring stability in the intelligent evaluation of music courses.

4.4 Performance comparison of different artificial intelligence evaluation models

This section compares the performance of MusicEval-Net with different AI evaluation models and observes the contribution of the model structure in combination with the ablation results. The comparison models include Audio-CNN, Behavior-LSTM, Text-BERT, and Late-FusionNet, which represent single Audio, single action, single Text, and Late fusion methods, respectively. The same training set, validation set and test set were used for all models, and the evaluation metrics were kept as Accuracy, F1-score, MAE and Cohen's κ. To ensure consistent comparison conditions, the same batch size and number of training rounds were used for each model, and the early stopping rules of the validation set were consistent.

To show the recognition of different levels, Fig. 7 presents the classification results of MusicEval-Net on four types of teaching quality levels using confusion matrix. The main diagonal proportions of the four categories of excellent, good, qualified and to be improved are 0.93, 0.91, 0.89 and 0.88, respectively, and the misjudgments are mainly concentrated between adjacent grades. The proportion of excellent samples misjudged as good was 0.06, and the proportion of samples to be improved misjudged as qualified was 0.09. There was no large-span grade misjudgment, indicating that the model output had good correspondence with the teacher's rating scale. The adjacent grade misjudgment is concentrated in the samples whose singing performance is close to the threshold, and the teacher review record also shows that there is a fuzzy scoring boundary in this kind of samples.

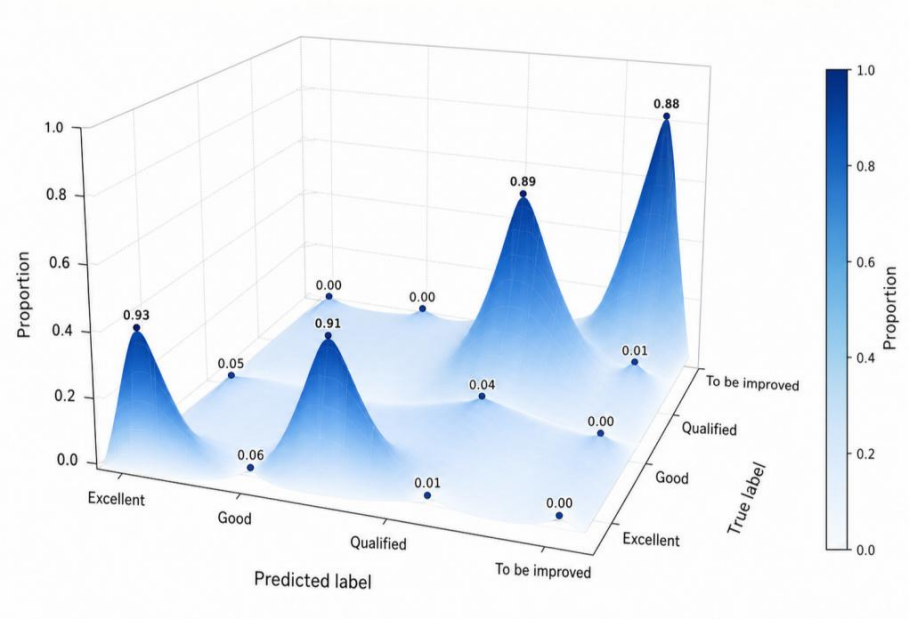

Figure 7: Confusion matrix for MusicEval-Net teaching quality level identification

To observe the distribution of scoring errors, Fig. 8 presents the absolute error density of each model using violin plots. The median error of MusicEval-Net is 0.18, and the interquartile range is 0.11. The error distribution of Audio-CNN and Behavior-LSTM is wider, and the tail samples are concentrated above 0.45. This result shows that multimodal fusion is more stable for complex classroom segments, and can especially alleviate the unimodal offset caused by recording noise or missing text feedback. Text-BERT performs better in feedback consistency samples, but the error expands significantly when facing vocal and rhythmic tasks, indicating that textual semantics is difficult to substitute for acoustic evidence.

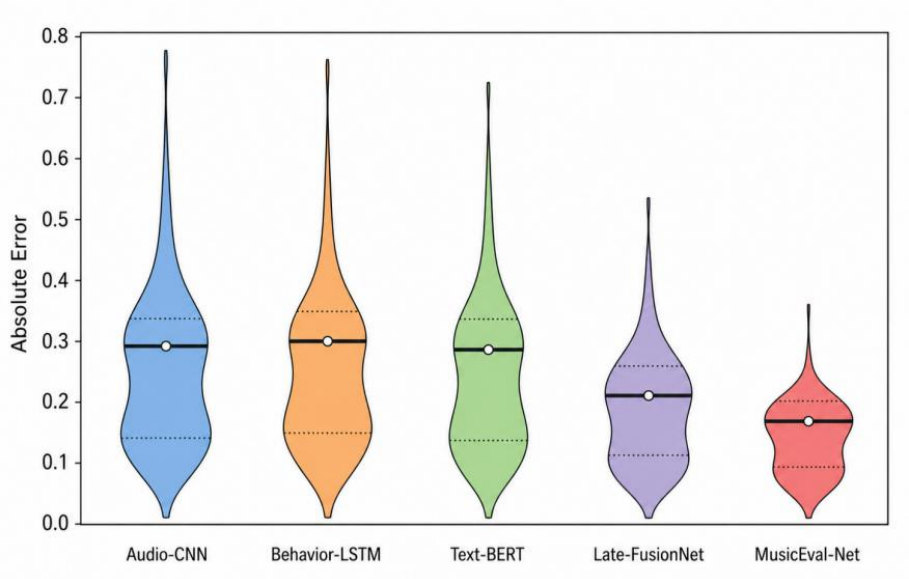

Figure 8: Violin plot of scoring error for different models

Table 4 presents the comprehensive performance of different models. MusicEval-Net achieves 91.8% Accuracy, 0.903 F1-score, 0.217 MAE and 0.84 Cohen's κ, which are better than the comparison models. Although Late-FusionNet uses multimodal input, it lacks task

attention and confidence calibration, resulting in an F1-score of 0.864 and an MAE of 0.286.

Table 4: Performance comparison of different AI evaluation models

Model	Accuracy / %	F1-score	MAE	Cohen's κ
Audio-CNN	84.6	0.821	0.341	0.69
Behavior-LSTM	83.2	0.807	0.333	0.67
Text-BERT	80.5	0.782	0.368	0.63
Late-FusionNet	86.9	0.864	0.286	0.76
MusicEval-Net	91.8	0.903	0.217	0.84

Taking the above together, it can be seen that the advantages of MusicEval-Net come from the combined effect of multi-modal feature alignment, task mapping, and feedback verification. The confusion matrix showed that the misjudgment of the model was mainly concentrated between adjacent grades, indicating that the scoring grade boundary was stable. The error distribution results show that the deviation of the model on most samples is small, indicating that the fusion structure can alleviate the impact of single modal data missing or fluctuation. Cohen's κ reached 0.84, indicating a high agreement between the model scores and the teacher review results. The results show that MusicEval-Net can meet the requirements of accuracy, stability and traceability of music teaching quality evaluation, and provide a reliable experimental basis for subsequent classroom terminal deployment.

4.5 Discussion

The results of MusicEval-Net in music course evaluation show that multimodal modeling is able to cover classroom evaluation evidence more completely than single acoustic identification. The experimental group had higher audio task completion rate, feedback viewing rate and classroom interaction frequency than the control group, indicating that the model output had entered the process of student practice, teacher review and result correction, rather than staying at score presentation. Correlation analysis showed that pitch stability, rhythm deviation, task completion rate and semantic activeness were related to vocal performance, learning participation and feedback consistency, respectively. The attention fusion layer can adjust the evidence weight according to the differences in vocal music, rhythm, instrumental music and appreciation tasks, so that the scoring basis is closer to the curriculum task. The model comparison results show that Audio-CNN, Behaviour-LSTM and Text-BERT can only cover local information. Although Late-FusionNet introduces multi-modal input, it lacks task mapping and confidence calibration, and the scoring error is still high. MusicEval-Net maintains good performance on Accuracy, F1-score, MAE and Cohen's κ, indicating that the model has stable results in scoring accuracy and manual consistency. Ablation experiments further show that audio features are the main basis for music performance evaluation, behavior logs can supplement the evidence of learning process, and text semantics can enhance feedback consistency judgment. Based on the above results, the intelligent evaluation system is more suitable as a computational aid for teacher grading. In practical applications, it is still necessary to retain the mechanisms of teacher review, low-confidence sample reflux and deviation detection, so that the algorithm output and professional judgment form the evaluation evidence chain together. The subsequent classroom deployment should also pay attention to the distribution shift caused by the differences in recording equipment, track difficulty and classroom tasks, and continuously record the collection conditions, model versions and review opinions on the system side to ensure that the evaluation results have the basis of stability, traceability and experimental reproduction.

5 Conclusion

Focusing on the teaching quality evaluation of music courses empowered by artificial intelligence, this paper constructs a MusicEval-Net multimodal intelligent evaluation model, which integrates classroom audio, behavior log and text feedback into a unified computing framework. The model organizes the evaluation evidence from three levels of acoustic features, learning process features and semantic feedback features, and forms technical links with the help of deep learning training, indicator mapping, teacher review and deviation detection. Research shows that multimodal fusion makes up for the shortcomings of single audio evaluation or single platform log analysis, and enables singing performance, rhythm training, practice engagement and feedback understanding in music classroom to obtain stable computational expression. The system set confidence calibration, result archiving and manual verification mechanisms to make the evaluation results back to the original sample, feature contribution and review record, and avoid the automatic scoring from the course context. The limitation of this paper is that the sample source is centralized, and the classroom type is mainly regular music courses, which has not yet covered complex scenes such as chorus rehearsal, instrumental ensemble, improvisation and cross-campus mixed classroom. Recording equipment, voice type, track difficulty and teacher's grading style may affect model generalization performance. Future research will expand the multimodal data set of music classroom and introduce more fine-grained acoustic labels, task labels and interaction labels to enhance the adaptation of the model to different music tasks. In the future, lightweight audio coding, self-supervised pre-training, edge-end inference and privacy-preserving learning are also studied to make the system operate stably in the real teaching platform. The intelligent evaluation results should retain the entrance of teacher review, and form a curriculum quality evaluation mechanism supported by algorithm judgment and professional judgment.

Major Project

2025 Beijing Higher Education Undergraduate Teaching Reform and Innovation Project: Innovative Practice of Digital and Intelligent-driven and AI-enabled Higher Music Education, Project Number: Pending

Acknowledgements

This work was supported by the China Conservatory of Music.

About the Author

Zijian Wang was born in Taiyuan, Shanxi, China, in 2000. He previously studied at Inner Mongolia Normal University and received a master's degree in 2025. Since 2025, he has become a doctoral student in the Music Technology Department of the China Conservatory of Music. He has published nine papers, one of which has been indexed by CSCI. His research fields include music technology, music acoustics and artificial intelligence.E-mail: 18335163369@163.com

Xiaodong Fu was born in Urumqi, Xinjiang, China, in 1972.He is currently a professor and a PhD candidate supervisor at the China Conservatory of Music. He is a ‘Changcheng’scholar

of Beijing and is also the director of the Academic Affairs Department and the librarian of the China Conservatory of Music. His main research interests include music technology and instrument acoustics.He has led over 20 research projects, published 3 monographs and one textbook, and has published over 40 papers, 8 of which have been indexed by CSSCI and CSCI. E-mail: 20200061@nuc.edu.cn

References

[1] Köktürk-Güzel B E, Büyük O, Bozkurt B, et al. Automatic assessment of student rhythmic pattern imitation performances[J]. Digital signal processing, 2023, 133: 103880.

[2] Faghih B, Chakraborty S, Yaseen A, et al. A new method for detecting onset and offset for singing in real-time and offline environments[J]. Applied Sciences, 2022, 12(15): 7391.

[3] Civit M, Civit-Masot J, Cuadrado F, et al. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends[J]. Expert Systems with Applications, 2022, 209: 118190.

[4] Ferreira P, Limongi R, Fávero L P. Generating music with data: Application of deep learning models for symbolic music composition[J]. Applied Sciences, 2023, 13(7): 4543.

[5] Paroiu R, Trausan-Matu S. Measurement of Music Aesthetics Using Deep Neural Networks and Dissonances[J]. Information, 2023, 14(7): 358.

[6] 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.

[7] Modran H A, Chamunorwa T, Ursuțiu D, et al. Using deep learning to recognize therapeutic effects of music based on emotions[J]. Sensors, 2023, 23(2): 986.

[8] Kang J, Poria S, Herremans D. Video2music: Suitable music generation from videos using an affective multimodal transformer model[J]. Expert Systems with Applications, 2024, 249: 123640.

[9] Kwiecień J, Skrzyński P, Chmiel W, et al. Technical, musical, and legal aspects of an AI-Aided algorithmic music production system[J]. Applied sciences, 2024, 14(9): 3541.

[10] Shahriar S. GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network[J]. Displays, 2022, 73: 102237.

[11] Martinez-Roig R, Cazorla M, Esteve Faubel J M. Social robotics in music education: A systematic review[C]//Frontiers in education. Frontiers Media SA, 2023, 8: 1164506.

[12] Kasneci E, Seßler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and individual differences, 2023, 103: 102274.

[13] Akgun S, Greenhow C. Artificial intelligence in education: Addressing ethical challenges in K-12 settings[J]. AI and Ethics, 2022, 2(3): 431-440.

[14] Hlosta M, Herodotou C, Papathoma T, et al. Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors[J]. Computers and Education: Artificial Intelligence, 2022, 3: 100108.

[15] Lamb R, Neumann K, Linder K A. Real-time prediction of science student learning outcomes using machine learning classification of hemodynamics during virtual reality and online learning sessions[J]. Computers and Education: Artificial Intelligence, 2022, 3: 100078.

[16] Laupichler M C, Aster A, Schirch J, et al. Artificial intelligence literacy in higher and adult education: A scoping literature review[J]. Computers and Education: Artificial Intelligence, 2022, 3: 100101.

[17] Caspari-Sadeghi S. Artificial intelligence in technology-enhanced assessment: A survey of machine learning[J]. Journal of Educational Technology Systems, 2023, 51(3): 372-386.

[18] Martinez-Comesana M, Rigueira-Díaz X, Larranaga-Janeiro A, et al. Impact of artificial intelligence on assessment methods in primary and secondary education: Systematic literature review[J]. Revista de Psicodidáctica (English ed.), 2023, 28(2): 93-103.

[19] Ouhaichi H, Spikol D, Vogel B. Research trends in multimodal learning analytics: A systematic mapping study[J]. Computers and Education: Artificial Intelligence, 2023, 4: 100136.

[20] Ashwin T S, Prakash V, Rajendran R. A systematic review of intelligent tutoring systems based on Gross body movement detected using computer vision[J]. Computers and education: Artificial intelligence, 2023, 4: 100125.

[21] Yildirim-Erbasli S N, Bulut O. Conversation-based assessment: A novel approach to boosting test-taking effort in digital formative assessment[J]. Computers and Education: Artificial Intelligence, 2023, 4: 100135.