



A Study on Emotional Intelligence Recognition and Classroom Teaching Adaptation Strategies in Belarusian National Music

Xuejun Zhai^{1,*}

¹ School of Music, Zhaoqing University, Zhao'qing 526000, Guangdong, China

SUMMARY: *In order to improve the accuracy of students' emotion intelligent recognition and the real-time performance of teaching adjustment in the Belarusian folk music classroom, this paper constructs a computational framework that fuses multimodal perception and classroom adaptation decision-making. Based on 72 real classroom records, 118 students and 14 representative works, a multi-source dataset covering acoustic, visual, behavioral and teaching context information was established, and a gated fusion recognition model was designed to realize the joint representation of emotional perception, emotional understanding, collaborative participation and regulatory stability. On this basis, the generation mechanism and real-time feedback process of classroom teaching adaptation strategy are further constructed. Experimental results show that the Accuracy of the proposed model reaches 91.62% and Macro-F1 reaches 90.84%, which are 3.25% and 3.20% higher than those of the single Transformer model respectively. The strategy matching rate of the system adaptation group was 91.8%, the average response time was 1.84 s, and the improvement of classroom participation reached 18.7%. The research shows that embedding artificial intelligence methods into Belarusian folk music teaching can provide interpretable and deployable technical paths for classroom emotion recognition and precise intervention.*

KEYWORDS: *Belarusian folk music; Intelligent emotion recognition; Multi-modal fusion; Teaching adaptation strategy*

1 Introduction

With the continuous development of intelligent perception, educational data mining and digital music analysis technology, the Chinese music classroom is gradually shifting from the experience-led teaching field to a compound teaching system that can be calculated, feedback and adjusted. Belarusian folk music has distinct characteristics in melody direction, rhythm organization, timbres level and cultural semantics. Its classroom teaching not only involves singing, playing and understanding of works, but also accompanies complex psychological processes such as emotional perception, emotional investment, interactive response and aesthetic judgment [1]. Traditional classroom mainly relies on teachers' immediate observation and teaching experience to grasp students' emotional state and teaching adaptation time. Although this method is context-sensitive, it is difficult to stably identify the subtle differences of students' emotional changes in the continuous classroom process, and it is not easy to form repeatable and verified data basis [2]. Especially in the teaching of folk music, students' understanding of the emotional connotation of works is often affected by cultural background, music experience and classroom participation. If there is no dynamic

*yangerjiji@163.com

<https://doi.org/10.65102/is2026687>

recognition of emotional intelligence state, teaching adjustment often lags behind real learning response.

The existing research has made some progress in the direction of music emotion recognition, educational emotion computing and digital music teaching. The deep learning model has shown strong capabilities in audio feature extraction, facial expression recognition, speech emotion judgment and behavior sequence modeling, and the multimodal fusion method also provides a new technical path for state recognition in complex classroom situations [3]. However, related studies mostly focus on general music emotion classification, online learning environment or western music teaching scenarios, and there is still a lack of targeted data organization and task modeling framework for Belarusian folk music classroom, which has regional culture, performance interaction and teaching scene [4]. At the same time, existing systems tend to pay more attention to "whether the recognition is accurate", and lack of discussion on how to transform the recognition results into executable teaching adaptation strategies, resulting in an obvious fault between emotion computing and classroom decision-making [5].

Based on this, this paper focuses on the problem of emotional intelligent recognition and teaching adaptation in the classroom of Belarusian folk music, and tries to construct a complete technical link from multi-source data collection, emotional representation learning to classroom feedback generation. Based on classroom video, speech, singing audio, learning behavior records and teacher annotations, a multimodal data system for teaching scenarios was established. At the model level, an intelligent emotion recognition method combining acoustic features, visual expression features and interactive behavior features was introduced to describe the dynamic states of students in the dimensions of understanding, engagement, collaboration and regulation. At the application level, the recognition results were mapped into teaching adaptation strategies such as rhythm adjustment, content switching, group intervention and feedback reinforcement, so as to improve the responsiveness and teaching pertinence of folk music classroom. This work does not regard emotion recognition as an additional function outside the classroom, but embeds it into the teaching operation process, so that the computational model, classroom observation and teaching decision-making form a closed-loop connection. By combining the cultural expression characteristics of folk music with the educational computing method, this paper hopes to provide a research idea for the Belarusian folk music classroom that takes into account both technical feasibility and teaching interpretation power, and also provide a more scene-adaptive reference for the intelligent transformation of folk music education.

2 Related Research

In recent years, the research on music affective computing, the cultivation of emotional ability in music education and classroom intelligent intervention has continued to advance, and the relevant results provide an important reference for the intelligent recognition and teaching adaptation of emotions in the classroom of Belarusian folk music. The existing research focuses on the construction of music emotion recognition models, focusing on the problems of audio feature representation, temporal dependence modeling and multi-modal fusion. Louro et al. compared the performance of various deep learning methods in music emotion recognition and pointed out that the combination of convolutional network and temporal network was more suitable for dealing with dynamic emotional changes in music clips [6]. Han et al. further improved the stability of emotion discrimination through the Inception-GRU residual structure [7]. Zhao et al. introduced a hierarchical cross-modal attention

mechanism to jointly model audio, text and other modal information, indicating that a single acoustic feature is difficult to fully support emotional interpretation in complex music context [8]. Such research provides a method basis for multimodal recognition in classroom environment, but its task goal mostly remains in "what emotion is expressed by music", and less touches on the educational problem of "how learners perceive, understand and regulate emotions in the classroom".

Another type of research focuses on the emotional dimension and digital teaching practice in music education. Liang, Varadi et al., Culp et al pointed out that emotional engagement, empathy ability and social interaction in music learning have become important variables to evaluate teaching quality [9-11]. Starting from the relationship between music empathy and learning engagement, Tu and Fu proved that emotional experience would directly affect learning persistence and subjective sense of gain [12]. At the same time, Ouyang, Maharaj and Gill, Liu and Shao and other studies show that mobile learning, online courses and digital platforms are reshaping the organization of music teaching, so that classroom behaviors, feedback rhythms and participation tracks can be recorded and analyzed in the form of data [13-15]. Feng's discussion on the reform of digital music education in China and Belarus also revealed the realistic demand for technology embedding in the context of Belarusian music education [16]. However, these researches focus more on teaching improvement supported by technology, and less on constructing intelligent emotion recognition links oriented to specific classroom situations. In particular, there is a lack of system design that directly maps the recognition results into teaching adaptation strategies.

From the existing results, related research has formed a rich accumulation in music emotion recognition, music education emotion development and digital teaching support respectively, but there are still obvious gaps between the three. On the one hand, the computational model emphasizes on the music object itself, and lacks a continuous representation of the emotional intelligence state of classroom participants. On the other hand, although educational research emphasizes emotional value, it generally lacks deployable recognition models and real-time feedback mechanisms. For the classroom of Belarusian folk music, there is a stronger coupling between the cultural semantics of works, the way of singing and playing, the atmosphere of group interaction and students' emotional reactions. If the general framework of music emotion classification is still used, it is often difficult to explain the real classroom differences in the teaching of folk music. Based on this, this paper attempts to combine multimodal emotion computing, classroom behavior analysis and teaching adaptation decision-making to construct an intelligent emotion recognition and strategy generation framework for Belarusian folk music classroom. The relevant research vein is shown in Table 1.

Table 1: Related research types, representative contents and implications for this paper

Research Direction	Representative References	Main Content	Existing Limitations	Implications for This Study
Music Emotion Recognition	[1][2][3][4]	Uses deep learning, multimodal fusion, and temporal modeling to recognize musical emotions	Most studies focus on music segments themselves and lack modeling of learners' classroom states	Acoustic feature extraction and cross-modal fusion methods can be adopted
Emotional Development in Music Education	[5][6][7][8][20]	Discusses the relationship between emotional ability, empathic engagement, and music learning outcomes	Strong in educational interpretation, but insufficient in computational implementation	Provides a theoretical basis for designing emotional intelligence dimensions
Digital and Mobile Music Teaching	[9][10][11][13][14][18]	Emphasizes the influence of platformization, mobility, and technological intervention on the teaching process	Few studies establish a real-time recognition–feedback closed loop	Supports the design of classroom behavioral data collection and feedback mechanisms
Music Teaching in Cultural Contexts	[12][15][16][17][19]	Focuses on ethnic music, cross-cultural teaching, and data-driven emotional support	Lacks a systematic model tailored to Belarusian national music classrooms	Demonstrates the necessity of this study in terms of object selection and scenario modeling

3 Intelligent emotion recognition and teaching adaptation method for Belarusian folk music classroom

3.1 Belarussian folk music classroom situation and task definition

The intelligent emotion recognition in Belarusian folk music classroom is not an isolated judgment of students' facial emotions or classroom atmosphere, but a continuous computing task centered on classroom perception, music understanding, collaborative participation and self-regulation. Compared with general music teaching, this kind of classroom includes various activities such as melody miming, rhythm following, semantic understanding of lyrics, cultural background understanding, group training and teachers' real-time demonstration. Students' emotional state often changes rapidly between listening, responding, imitation, discussion and correction, and its change is not only affected by pitch, strength, speed and

timeliness. It is also closely related to the difficulty of classroom tasks, the intensity of peer interaction and the strangeness of folk music culture. Traditional classroom mainly relies on teacher observation, after-class interview or questionnaire results to infer students' emotional responses. This method can provide certain experience judgment, but it is difficult to stably depict the fine-grained fluctuations of students' emotional intelligence in the continuous teaching process, and it is more difficult to form reusable data support. Based on this, this paper models the Belarusian folk music classroom as a dynamic interaction field composed of teaching event flow, student behavior flow and emotional feedback flow, whose task structure is shown in Figure 1.

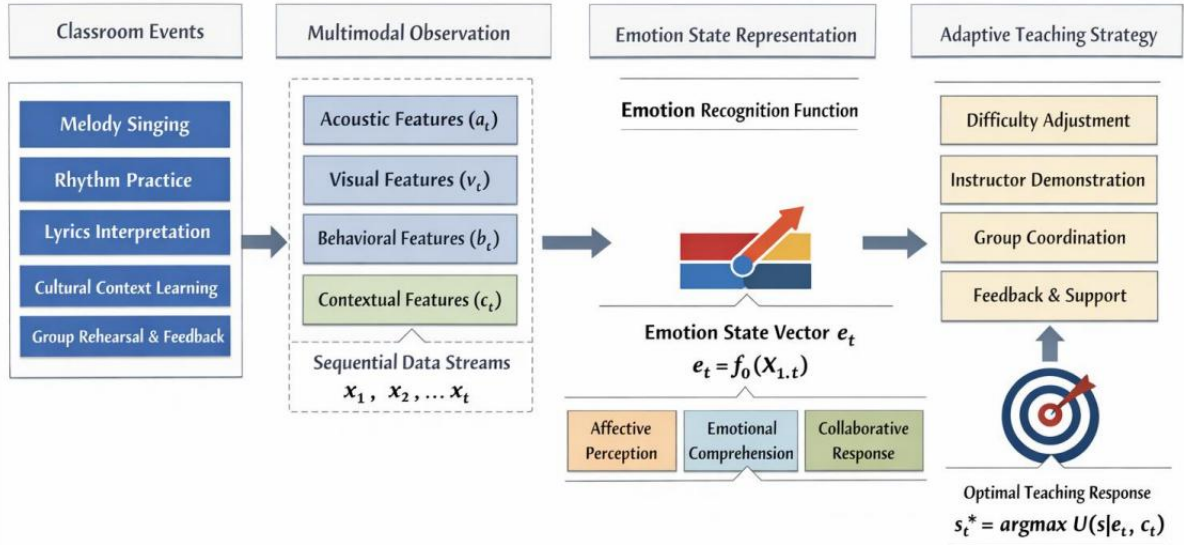


Figure 1: Task structure framework oriented to Belarusian folk music classroom

In this study, emotional intelligence is not simplified to the static labels of "positive-negative" or "pleasure-down", but is defined as the students' ability to perceive the emotional cues of music in a specific classroom situation, the ability to understand the emotional connotation of the work, the collaborative response ability in group activities, and the ability to regulate when frustration or deviation appears. Let the observation vector at the t -th class moment be:

$$x_t = [a_t; v_t; b_t; c_t] \quad (1)$$

Among them, a_t represents the acoustic features composed of singing audio, speech clips, and respiratory rhythms, v_t represents visual features such as facial expressions, gaze direction, and body posture, b_t represents behavioral features such as clicking, following, response rounds, and interaction delays, and c_t represents situational features such as work paragraphs, teaching activity types, rhythm structures, and teacher instructions. For the continuous window $X_{1:T} = \{x_1, x_2, \dots, x_T\}$, the emotional intelligence state in the classroom can be expressed as:

$$e_t = f_\theta(X_{1:t}) \quad (2)$$

where e_t is the intelligent representation vector of emotion at time t , and f_θ is the intelligent recognition function of classroom emotion. The significance of this expression is that it no longer treats students' classroom responses as scattered signals, but maps multi-source data

into a learnable state space, which provides a computational basis for subsequent teaching adaptation.

The complexity of the Belarusian folk music classroom is also reflected in the fact that similar representations may have different teaching implications in different classroom sessions. For example, in the stage of listening to lyrical folk songs, singing softly, focusing eyes and converging movements may indicate a higher level of emotional immersion. Similar performance during rhythm training or phonation sessions could also mean slow responses, difficulty following beats, or decreased engagement. Therefore, in this paper, the task is further defined as a sequence discrimination problem subject to context constraints, whose context mapping relationship is shown in Figure 2.

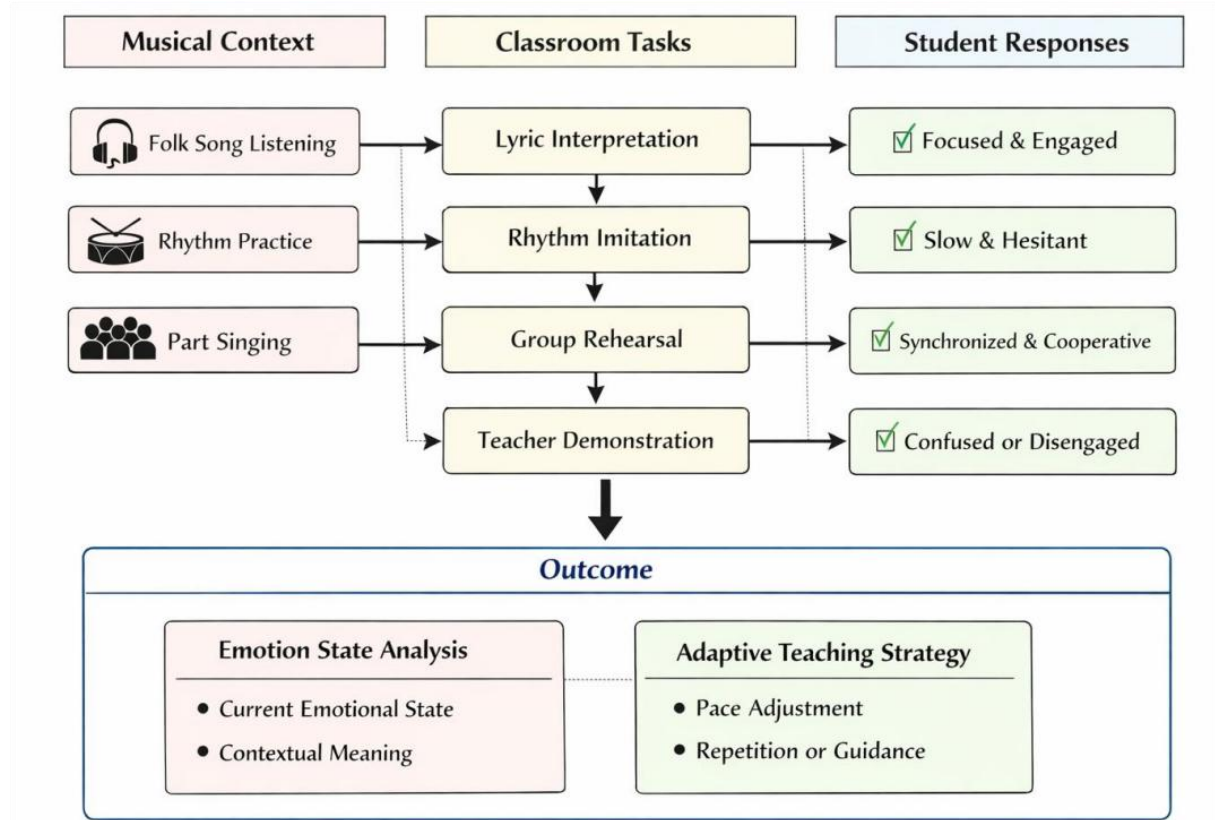


Figure 2: Coupling diagram of "situation-reaction-strategy" in Belarusian folk music classroom

In order to use the recognition results for classroom adjustment, this paper defines the teaching adaptation task as a strategy optimization problem based on the recognition results. Let the set of classroom strategies be S , then the optimal instructional response at time t is written as:

$$s_t^* = \arg \max_{s \in S} U(s | e_t, c_t) \quad (3)$$

where $U(s | e_t, c_t)$ represents the adaptation utility of strategy s under the condition of emotional intelligence state e_t and classroom situation c_t , and s_t^* is the optimal teaching strategy output by the system. This strategy is not an abstract suggestion, but an executable operation oriented to the real classroom, such as slowing down the rhythm, replaying the teacher's demonstration, strengthening the group collaboration, supplementing the semantic of lyrics, reducing the difficulty gradient, or enhancing the emotional cue. Therefore, the

association between intelligent emotion recognition and instructional adaptation is no longer loose, but forms a continuous closed loop from state perception, meaning judgment to strategy generation.

In the specific modeling, this paper divides the Belarusian folk music classroom into five task units: introduction perception, melody learning, lyrics understanding, coordination and feedback correction. The observation focus and calculation goal of different units are different. See Table 2 for the relevant definitions. Through this task decomposition, the classroom is no longer regarded as a single teaching process, but is organized as a number of computational fragments with clear inputs, states and outputs. The significance of this process is that, on the one hand, the cultural characteristics and teaching authenticity of the folk music classroom are retained, on the other hand, the multi-modal recognition model can learn the emotional intelligence change rules in different links under a unified framework, which provides a consistent problem expression basis for subsequent data collection, model construction and real-time feedback process design.

Table 2: Definition of task units and computational objectives for Belarusian folk music classroom

Task Unit	Main Teaching Activity	Key Observed Variables	Main Computational Objective
Introductory Perception	Listening to the musical work and forming an initial perception of its emotional tone	Gaze concentration, facial arousal level, and initial vocal response	Emotional perception state recognition
Melody Learning	Melody imitation and rhythm following	Pitch deviation, rhythm synchronization, and posture stability	Assessment of engagement and following ability
Lyric Understanding	Analysis of textual content and cultural semantics	Verbal responses, pause duration, and facial expression changes	Estimation of the depth of emotional understanding
Ensemble Collaboration	Group singing and interactive coordination	Part-singing coordination, response latency, and peer gaze behavior	Recognition of collaborative participation level
Feedback-based Correction	Teacher comments and students' secondary adjustment	Correction magnitude, recovery speed, and repeated error rate	Evaluation of regulation ability and adaptation effectiveness

3.2 Multi-source classroom data collection and annotation system construction

In order to provide a stable input basis for the intelligent recognition of emotions in the classroom of Belarusian folk music, this paper further constructs a multi-source classroom data system for real teaching scenarios based on the above task definition. Different from general music emotion recognition, which only deals with audio clips, this study is faced with a continuous process of "work interpretation, teacher guidance, student response and classroom regulation". If the data only retain a single acoustic channel, it is difficult to explain the formation mechanism of students' emotion understanding, participation and regulation.

Based on this, this paper collected 72 classroom recordings of Belarusian folk music, covering 118 students, 6 teachers and 14 representative Belarusian folk music works, forming the original teaching data with a total duration of about 86.4 hours. Data sources include classroom panoramic videos, close-range facial videos, environmental and collar microphone audio, singing task results, classroom interaction logs, and teacher observation records, and the organization of each source data is shown in Table 3.

Table 3: Composition of multi-source data for Belarusian folk music classroom

Data Source	Collection Method	Main Parameters	Main Purpose
Panoramic Classroom Video	Fixed-camera recording	25 fps, 1920×1080	Records classroom organization, group interaction, and teacher regulation behaviors
Close-up Facial Video	Front-row targeted capture	25 fps, local cropped view	Extracts facial expressions, gaze, and head posture changes
Classroom Audio	Ambient microphone + lavalier microphone	22.05 kHz, mono	Extracts singing, response, and speech emotion cues
Behavioral Logs	System event recording	Millisecond-level timestamps	Records clicks, responses, singing-following latency, and correction frequency
Teacher Observation Records	In-class annotation and post-class review	Structured text	Assists in confirming emotional intelligence states and teaching intervention types

In order to ensure the temporal consistency of subsequent model training, all devices were synchronized with a unified timestamp, the video frame rate was set to 25 fps, the original audio was uniformly resampled to 22.05 kHz, interactive events were recorded in the form of millisecond log, and the continuous classroom flow was segmented into computable segments according to a 12 s sliding window. The processing flow is shown in Figure 3.

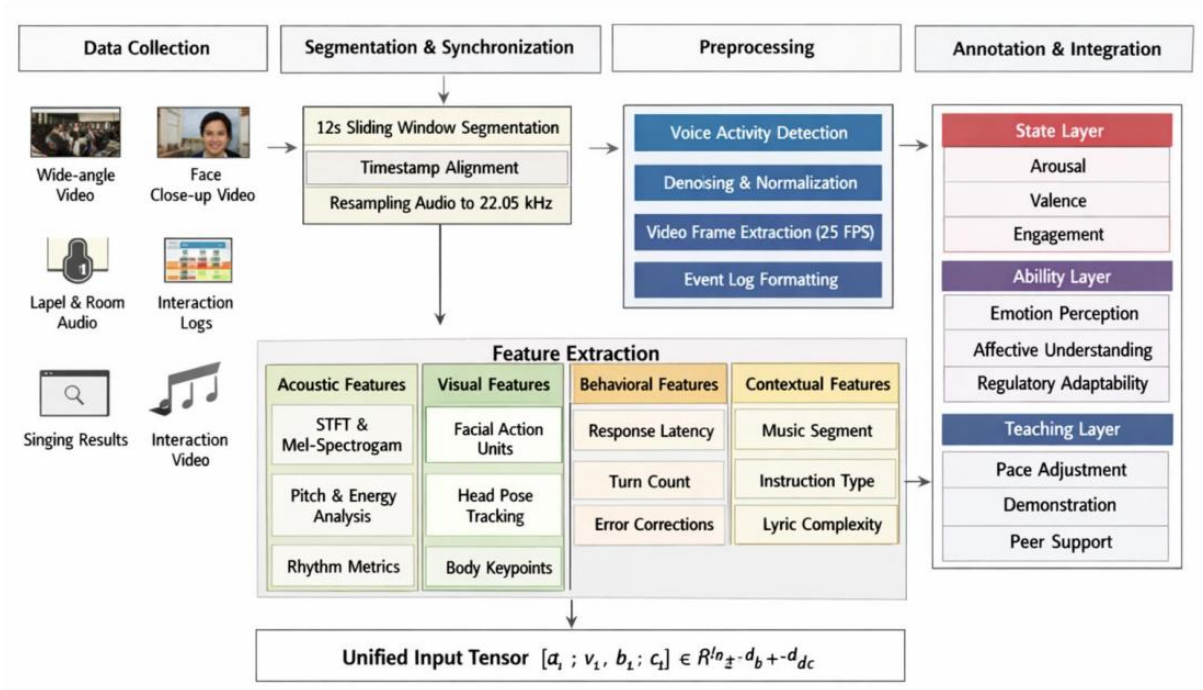


Figure 3: Process of multi-source classroom data collection and annotation

In the preprocessing stage, this paper establishes a unified coding method around four types of information: acoustic, visual, behavioral and situational. For class singing and speech signals, endpoint detection, noise reduction and loudness correction are performed first, and then time-frequency distribution is extracted by short-time Fourier transform:

$$S(\tau, \omega) = \left| \sum_{n=0}^{N-1} x(n)w(n - \tau)e^{-j\omega n} \right|^2 \quad (4)$$

Here, $x(n)$ is the original audio sequence, $w(\cdot)$ is the window function, τ represents the time position, and ω represents the frequency index. On this basis, the features of Mel spectrum, fundamental frequency trajectory, energy envelope, rhythm stability and speech rate change are further generated to describe the emotional expression intensity and sound control state of students in folk music singing. The visual channel extracts the dynamic changes of expression action units, gaze offset, head posture and 17 key points from the facial region and the upper body skeleton to represent the attention concentration, interaction intention and body cooperation. The behavioral channel recorded the start time delay of students singing along, response rounds, error correction times, speech length and group collaboration frequency. The situational channel encoded the current work paragraph, teaching link, rhythm type, semantic difficulty of lyrics and teacher intervention. All continuous features are normalized by Z-score:

$$\hat{x} = \frac{x - \mu}{\sigma} \quad (5)$$

To weaken perturbations of distributions by individual differences, recording environments, and class batches.

In the annotation system design, this paper does not follow the single emotion category label, but constructs a three-layer annotation structure of "state layer, ability layer and

teaching layer". The state layer marks the emotional arousal level, emotional tendency and participation intensity of students in the segment. The ability layer corresponds to four dimensions of emotion perception, emotion understanding, collaborative response and regulatory stability. The teaching layer records whether teachers implement adaptation measures such as slow pace, demonstration replay, layered questioning, peer assistance and immediate encouragement, the main contents of which are shown in Table 4.

Table 4: Design of intelligent annotation system for emotion

Annotation Level	Core Dimension	Label Description	Output Form
State Level	Arousal level, affective tendency, and participation intensity	Describes immediate responses within a segment	Three-level or five-level discrete labels
Ability Level	Perception, understanding, collaboration, and regulation	Describes the student's emotional intelligence ability state	Four-dimensional score vector
Teaching Level	Rhythm adjustment, demonstration replay, tiered questioning, encouraging feedback, etc.	Describes teacher adaptation measures	Multi-label event encoding

The labeling work was completed by two music education researchers and two teachers with classroom experience, and the method of "machine screening + manual review" was used to generate labels. The system first gives the candidate intervals according to acoustic energy, expression changes and behavior abnormalities, and then the candidate segments are manually corrected segment by segment to improve the labeling efficiency of long-term classroom data. The consistency test was carried out on 20% of the samples before the experiment, and Cohen's Kappa reached 0.84, indicating that the labeling system had good stability and reusability. After cleaning and labeling, this paper organizes each class segment into a unified input tensor:

$$X_t = [a_t; v_t; b_t; c_t] \in \mathbb{R}^{d_a+d_v+d_b+d_c} \quad (6)$$

Here, a_t, v_t, b_t and c_t represent acoustic, visual, behavioral and situational feature sub-vectors, respectively. All the samples were loaded into the library in a hybrid way of "structured table +JSON extension", which not only retained the original timestamp, fragment path and label index, but also supported the parallel loading and batch training of subsequent multimodal models. Through this data construction process, the emotional intelligence states in the Belarusian folk music classroom were transcribed into digital objects with time continuity, semantic hierarchy, and teaching interpretability, which provided a stable data base for subsequent recognition model design and teaching adaptation strategy generation.

3.3 Design of multimodal emotion intelligent recognition model

In order to improve the stability and discrimination accuracy of emotion intelligent recognition in Belarusian folk music classroom, this paper constructs a multi-modal fusion recognition model for teaching scenarios, and completes the joint modeling of acoustic information, visual behavior, classroom interaction and situational semantics under a unified framework. The model does not regard students' classroom responses as discrete signals, but maps the emotional ups and down, facial expression and posture changes, interactive behavior

rhythm and teaching context in the singing audio into a unified representation space, so as to realize the continuous recognition of four dimensions of emotional perception, emotional understanding, collaborative participation and regulatory stability. Its overall structure is shown in Figure 4.

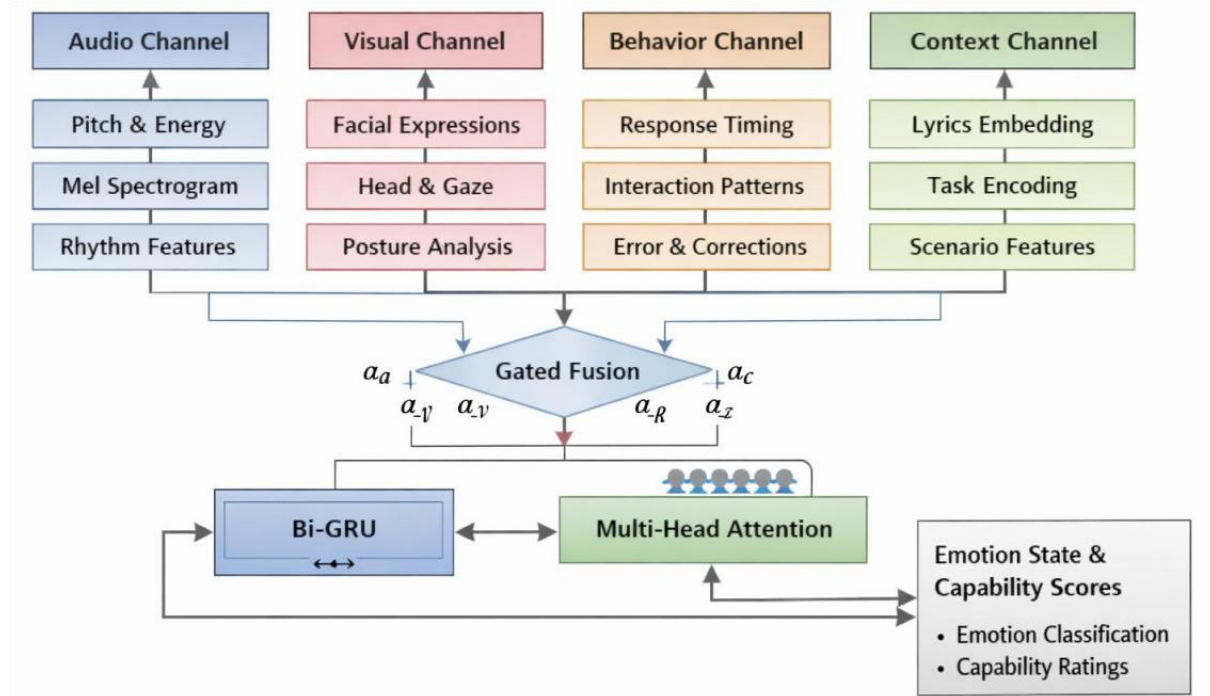


Figure 4: Structure diagram of multimodal emotion intelligent recognition model

In the feature fusion stage, the gated weighting mechanism is used instead of the simple splicing method to reduce the interference caused by the inconsistency of different modal noise levels. Let the outputs of the four types of subchannels at time t be $h_t^a, h_t^v, h_t^b, h_t^c$ respectively, then the fused representation is defined as:

$$z_t = \alpha_t^a h_t^a + \alpha_t^v h_t^v + \alpha_t^b h_t^b + \alpha_t^c h_t^c \quad (7)$$

Here, α_t^m is the adaptive weight of mode m at the current time and is satisfied:

$$\sum_{m \in \{a,v,b,c\}} \alpha_t^m = 1 \quad (8)$$

This design can dynamically adjust the importance of different information sources according to the classroom state, such as increasing the weight of behavioral channel and acoustic channel in the training stage, and enhancing the contribution of visual and situational channels in the lyrics understanding stage. In order to further preserve the temporal dependence, the fusion vector z_t is sent to the bidirectional gated recurrent unit and the multi-head attention module to extract the state continuation relationship and key reaction nodes across segments, and finally output the emotional intelligent representation vector e_t . The prediction results of the model not only include the immediate state label, but also output a four-dimensional capability score, which is calculated as:

$$\hat{y}_t = \text{Softmax}(W_o e_t + b_o) \quad (9)$$

Here, \hat{y}_t represents the probability distribution of the class segment in the target label space. Considering the imbalance of complex states such as "high engagement but insufficient understanding" or "stable emotion but low participation" in the classroom samples, this paper uses the weighted cross entropy and mean square error joint loss function in the training stage to optimize the classification results and ability scores simultaneously:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{mse} \quad (10)$$

Among them, \mathcal{L}_{ce} is used to constrain discrete state recognition, \mathcal{L}_{mse} is used to fit emotional intelligence four-dimensional score, and λ_1 and λ_2 are balance coefficients. In the model training, the acoustic and visual subnetworks used hierarchical Dropout and normalization strategies to suppress overfitting, and the behavior and context channels controlled the parameter scale by embedding compression. The key configurations are shown in Table 5.

Table 5: Key configurations of multimodal emotion intelligent recognition model

Module	Input Content	Core Structure	Output Function
Acoustic Channel	Mel spectrogram, fundamental frequency, energy, and rhythm stability	CNN + Temporal Convolution	Extracts emotional expression and vocal control features
Visual Channel	Facial action units, gaze, and head-shoulder posture	CNN + BiGRU	Captures attention state and interaction intention
Behavioral Channel	Response latency, number of corrections, and collaboration frequency	MLP + GRU	Describes classroom participation rhythm
Contextual Channel	Musical passage, task type, and teacher intervention mode	Embedding + Self-Attention	Provides semantic constraints and scene interpretation
Fusion and Output Layer	Four-channel hidden vectors	Gated weighting + multi-head attention + Softmax	Outputs state labels and ability scores

Through the structure design of "sub-channel modeling-gated fusion-timing discrimination-multi-task output", the multi-source heterogeneous signals in the classroom were organized into emotional intelligent representations with unified semantics, which provided an interpretable and real-time state basis for the generation of subsequent teaching adaptation strategies.

3.4 Generation mechanism of classroom teaching adaptation strategy

In order to realize the effective transformation from the results of intelligent emotion recognition to classroom intervention actions, this paper designs a teaching adaptation generation mechanism including state mapping, strategy screening, belief control and result writeback after the multimodal recognition model, so that the system output does not stop at the emotional judgment layer, but can further serve the real-time teaching adjustment in the classroom of Belarusian folk music. Different from general recommendation teaching support, this paper is not facing a static learning resource distribution task, but a continuously evolving classroom event flow. Therefore, the strategy generation must consider the coupling

relationship between students' current emotional intelligence state, teaching links and teachers' executable actions. Let the emotional intelligence representation output by the recognition model at time t be e_t and the classroom situation vector be c_t , then the state discrimination result can be written as:

$$p_t = \text{Softmax}(W_p[e_t; c_t] + b_p) \quad (11)$$

Among them, p_t represents the probability distribution of the current class segment on the target state space, which not only reflects the students' immediate participation degree, but also reflects their understanding and adjustment level of the emotional connotation of the work. Considering that there are often composite states such as "high investment but insufficient understanding" or "stable emotion but weak collaboration" in the classroom of Belarusian ethnic music, this paper does not adopt a rigid mechanism that a single label directly triggers a single action, but establishes a candidate strategy set \mathcal{A}_t , and retain several intervention programs most relevant to the current state, and then combines with the classroom context to complete the sorting and screening.

In the design of policy library, this paper constructs a three-layer adaptation structure. The basic response level was to instant classroom control, which mainly included rhythm slowdown, paragraph replay, pitch cue, semantic supplement of lyrics and teacher demonstration enhancement. The organizational adjustment level to the group interaction process, including group reorganization, round adjustment, collaborative practice strengthening and peer singing guidance; The individualized intervention level corresponded to the differentiated support for specific students or groups, such as the implementation of low-pressure prompts for students with high stress, the supplement of cultural background explanations for students with delayed understanding, and the increase of low-threshold response opportunities for students with insufficient participation. In order to compare the adaptability of actions at different levels in a unified framework, this paper defines a policy utility function:

$$U(a_i|e_t, c_t) = \lambda_1\phi(e_t, a_i) + \lambda_2\psi(c_t, a_i) + \lambda_3\eta(h_t, a_i) \quad (12)$$

where $\phi(e_t, a_i)$ represents the matching degree between strategy a_i and the current emotional intelligence state, $\psi(c_t, a_i)$ represents the context consistency between strategy a_i and work paragraphs, teaching task types, and classroom rhythm, $\eta(h_t, a_i)$ represents the effectiveness estimation of strategy under history classroom feedback record h_t , and $\lambda_1, \lambda_2, \lambda_3$ are the weight coefficients. The system thus outputs the optimal classroom response:

$$a_t^* = \arg \max_{a_i \in \mathcal{A}_t} U(a_i|e_t, c_t) \quad (13)$$

Thus, the recognition results are transcribed into teaching actions with clear execution implications. In order to avoid the excessive intervention of the model when the boundary of the state is fuzzy, this paper introduces the confidence filtering and candidate reservation mechanism. When $\max(p_t) \geq \tau$, the system directly outputs the optimal policy. When $\max(p_t) < \tau$ and the probabilities of multiple states are close, the Top-K candidate strategies are retained, and low-intensity intervention actions with less disturbance to classroom continuity are preferred, such as brief prompts, local sing along correction, or teacher's eye feedback, but high-intensity organizational adjustment is not immediately triggered. This design makes the strategy generation more robust and more in line with the teaching logic of "gradual correction" rather than "frequent interruption" in the real classroom. The system

output results are stored in the structured format of "timestamp -- object unit -- state judgment -- candidate strategy -- execution result", and written into the teacher's terminal interface synchronously for teachers to view and review after class. Figure 5 shows the overall process of the mechanism.

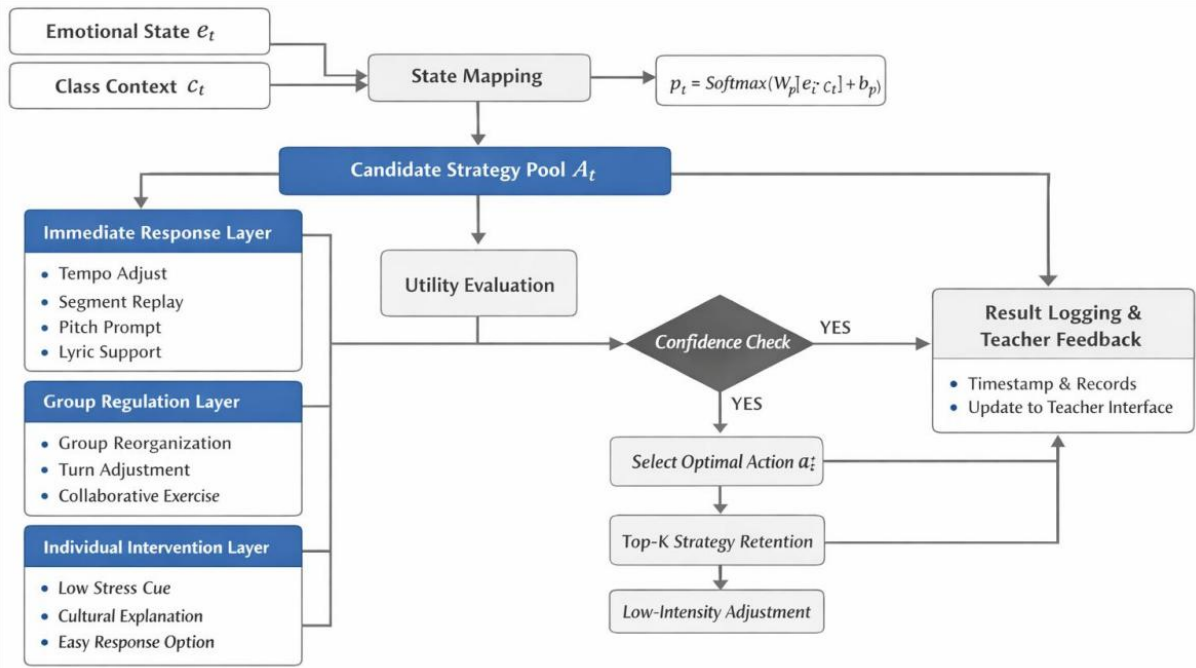


Figure 5: The mechanism framework of classroom teaching adaptation strategy generation

3.5 System implementation and real-time feedback process design

In order to make the emotion intelligent recognition and teaching adaptation mechanism truly enter the Belarusian folk music classroom rather than stay at the offline analysis level, this paper takes real-time, stability and interpretability as the core constraints in the system implementation stage, and uniformly designed the model calling process, data transmission link and teacher feedback method. Considering the continuous stream of audio and video in the classroom scene, the rapid change of students' responses, and the short decision window of teachers, if the system's reasoning delay is too high or the results are too complex, even if the recognition accuracy is high, it is difficult to support the immediate intervention in real teaching. Based on this, this paper adopts the lightweight deployment structure of "edge acquisition + local reasoning + interface writeback", connects the classroom camera, environmental microphone and interactive terminal to the unified acquisition layer, and sends them to the sliding window processing module after timestamp alignment, and then completes the state estimation by the compressed multi-modal recognition model, and finally pushes the adaptation strategy to the teacher terminal interface synchronously. Its process is shown in Figure 6. When the system runs, the system is continuously updated with an 8 s analysis window and a 2 s step, so that the classroom state not only retains the timing context, but also avoids weakening the feedback agility due to the long window.

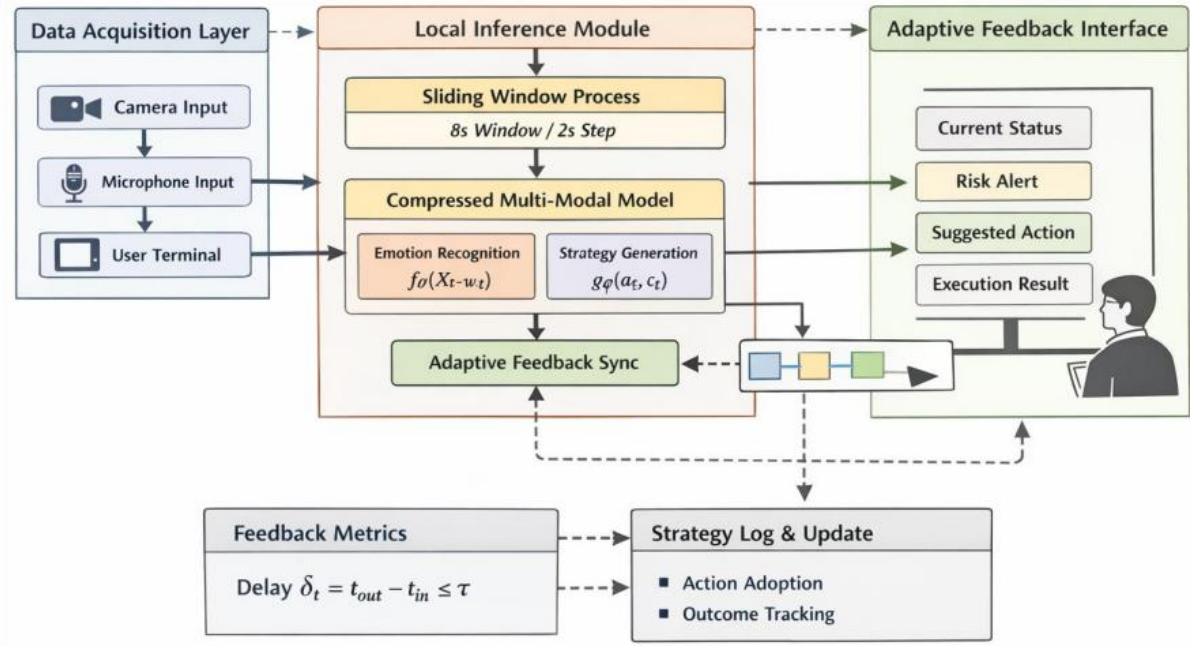


Figure 6: Flowchart of system implementation and real-time feedback

In terms of deployment and implementation, the original recognition model is converted into a lightweight reasoning version, the redundant convolutional blocks in the visual branch and the high-overhead attention layer in the temporal branch are trimmed, and 16-bit mixed-precision calculation is combined to reduce the video memory consumption and reasoning burden. For an input fragment $X_{t-w:t}$ at any time t , the system output can be written as:

$$\hat{a}_t = g_\phi(f_\theta(X_{t-w:t}), c_t) \quad (14)$$

where $f_\theta(\cdot)$ represents the emotion intelligent recognition model, $g_\phi(\cdot)$ represents the teaching adaptation strategy generation module, c_t is the current classroom situation information, and \hat{a}_t is the output real-time teaching suggestions. In order to measure whether the system meets the requirements of classroom application, this paper introduces the feedback delay index:

$$\delta_t = t_t^{\text{out}} - t_t^{\text{in}} \quad (15)$$

Here, t_t^{in} represents the moment when the window data enters the inference queue, and t_t^{out} represents the moment when the teacher side receives the suggestion. The system design objective is to make $\delta_t \leq \tau$, where τ is the acceptable classroom feedback threshold. In order to reduce the result jitter caused by sudden fluctuations, a short-term smoothing mechanism is added to the output layer, and high-intensity intervention suggestions are triggered only when similar state judgments occur in two consecutive update cycles. If the change of the state probability is small, the interface only displays low-intensity prompt information, such as "slow down the pace" or "supplement the demonstration", to maintain the continuity of the class.

4 Experimental design and result analysis

4.1 Data set composition and experimental setup

In order to verify the effectiveness of the intelligent emotion recognition and classroom teaching adaptation method constructed in this paper in the teaching scene of Belarusian folk music, the experiment part is carried out based on the multi-source classroom data system established in the previous section. The whole data set comes from 72 real classroom recordings, covering 118 students, 6 teachers and 14 representative Belarusian folk music works. The total duration of the original data is about 86.4 hours. Considering the continuous evolution characteristics of classroom emotional states, this paper first annotates the original clips with a 12 s semantic window, and then reconstructs the training samples according to the input length of 8 s and the sliding step size of 2 s, so as to balance the retention of classroom context and the efficiency of model calculation. After segmentation and denoising, a total of 18436 effective sample units were formed, in which each sample included acoustic features, visual behavior features, interaction log features and teaching situation labels at the same time, and corresponding to emotional intelligence state labels and teaching adaptation records. In order to reduce the risk of data leakage caused by individual repetition, this paper divides the batch by joint constraint of student and class, and generates the training set, validation set and test set according to the 8:1:1 ratio to ensure that the same class session is not distributed across sets.

The experimental platform is deployed in Linux environment, the backend is implemented with Python 3.11 and PyTorch 2.2, the GPU is NVIDIA A100 40 GB, and the CUDA version is 12.1. AdamW optimizer was used in the model training process, the initial learning rate was set to 2×10^{-4} , the batch size was set to 32, and the maximum training round was 40. If the validation set index did not improve for 5 consecutive rounds, Early Stopping was triggered to prevent overfitting. All multimodal inputs perform uniform normalization before entering the model, where audio is kept at 22.05 kHz sampling rate, video frame rate is fixed at 25 fps, and behavior logs are mapped to a uniform window after alignment by timestamp. Accuracy, Macro-F1 and weighted Kappa were used to measure the consistency of category discrimination in the recognition task. The matching rate, response delay and correction effect after adoption were counted in the strategy generation part. In order to ensure the reproducibility of the experimental results, the random seeds, parameter configuration and log files were recorded in all training processes, and the average value was taken by repeating three times under the same partition conditions. The main situation of the dataset and the experimental setup is shown in Table 6. Through the above arrangement, the experiment can not only test the recognition ability of the model in multimodal classroom situations, but also further investigate its stability and practical adaptation value as a teaching support tool.

Table 6: Data set composition and experimental setup

Item	Content
Number of Original Classroom Sessions	72
Number of Students	118
Number of Teachers	6
Number of Musical Works	14 Belarusian national music works
Total Duration of Raw Data	86.4 h
Sample Construction Method	12 s annotation window, 8 s input segment, stride of 2 s
Number of Valid Samples	18,436
Data Split	Training/Validation/Test = 8:1:1
Input Modalities	Acoustic, visual, behavioral, and contextual
Runtime Environment	Linux + Python 3.11 + PyTorch 2.2 + CUDA 12.1
Hardware Configuration	NVIDIA A100 40 GB
Optimizer	AdamW
Initial Learning Rate	2×10^{-4}
Batch Size	32
Maximum Number of Training Epochs	40
Evaluation Metrics	Accuracy, Macro-F1, Weighted Kappa, strategy matching rate, and response latency

4.2 Performance analysis of intelligent emotion recognition

In order to verify the recognition ability of the model in the Belarusian ethnic music classroom, this paper compares the performance of acoustic single-channel model, visual single-channel model, behavioral single-channel model, situational single-channel model and multimodal fusion model on the test set, and uses Accuracy, Macro-F1 and weighted Kappa as the core evaluation indicators. Each model is trained under the same data partition and training environment to ensure the comparability of results. The experimental results show that although the single-channel model can capture the local cues of classroom emotional intelligence, it still has obvious limitations in complex teaching situations. Among them, the overall performance of the visual model is better than that of the other single-channel methods, with an Accuracy of 85.73%, Macro-F1 of 84.68%, and weighted Kappa of 0.823, indicating that facial expression, gaze and posture information have strong explanatory power for classroom emotional states. The Accuracy of acoustic model is 84.26%, and Macro-F1 is 83.41%, which indicates that the pitch fluctuation, energy change and rhythm control in students 'singing and responding can indeed reflect their emotional engagement level. In contrast, the independent performance of the behavior model and the situation model is weak, with the Accuracy of 81.94% and 79.88% respectively, indicating that relying solely on the classroom operation trajectory or the context of teaching fragments is not enough to stably depict the emotional intelligence state of students.

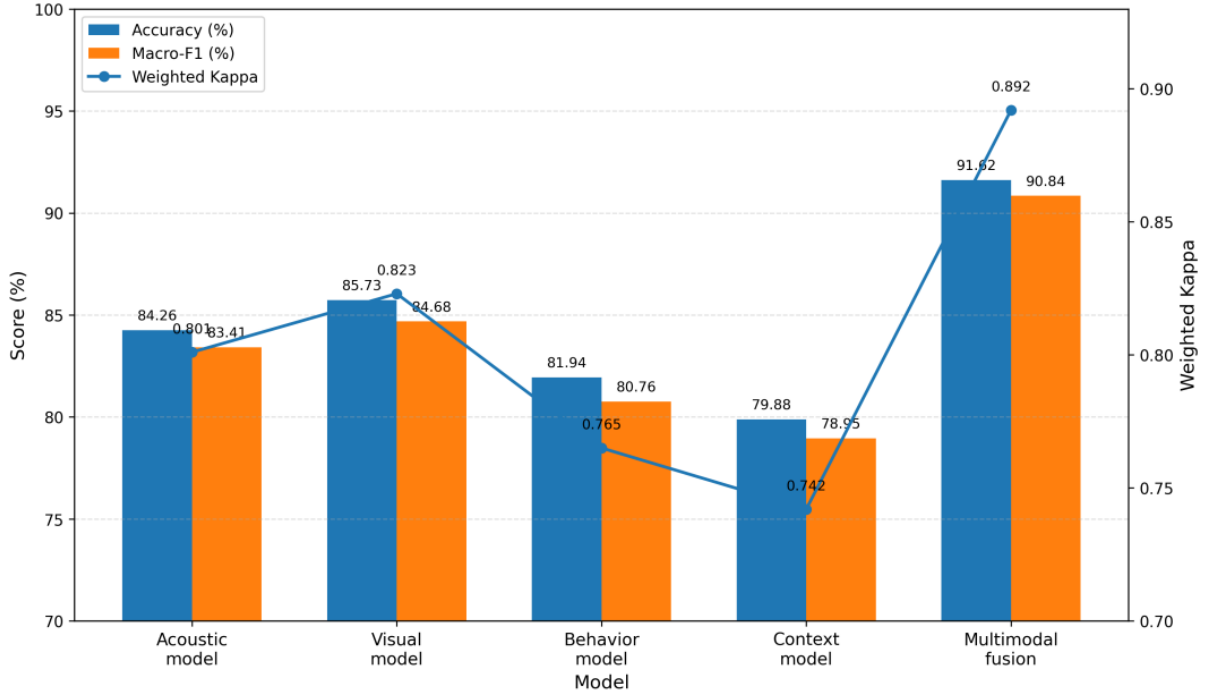


Figure 7: Performance comparison of different models on the task of intelligent emotion recognition

As shown in Figure 7, the multimodal fusion model achieves the best results on the three indicators, with an Accuracy of 91.62%, a Macro-F1 of 90.84%, and a weighted Kappa of 0.892, which are 5.89, 6.16, and 0.069 higher than those of the best single-channel model, respectively. This result shows that the emotional intelligence in Belarusian folk music classroom is not determined by a single signal, but a composite state composed of sound expression, expression posture, behavioral response and classroom context. The fusion model integrates the complementary information in different modalities into a unified representation space through gated weighting and time series modeling, thus effectively reducing the accumulation of bias in single channel judgment. From the boundary distribution of the recognition results, the discrimination ability of the fusion model in the adjacent states is improved, indicating that it not only improves the overall accuracy, but also improves the recognition stability of the boundary samples. The above results verify the effectiveness of the multimodal fusion structure in the intelligent recognition task of classroom emotions, and also provide a state input with higher credibility for the generation of subsequent teaching adaptation strategies.

4.3 Analysis of classroom teaching adaptation effect

In order to test the actual role of the teaching adaptation mechanism proposed in this paper in the classroom of Belarusian folk music, this paper further carried out a real teaching deployment test, and compared the system adaptation group, the teacher experience adjustment group and the fixed process teaching group as control objects. The test covered a total of 36 classroom activities, 62 students, and lasted for 4 weeks. The experimental platform adopts the real-time recognition and feedback system constructed in the previous section. The data of strategy triggering, teacher adoption, student response and after-class evaluation are synchronously recorded in the classroom process, and the matching degree of classroom rhythm, the support degree of teaching understanding and the emotional

participation experience are scored by Likert 5-level scale. The evaluation focuses not only on whether the system can give suggestions, but also on whether these suggestions can be translated into visible teaching improvement in successive classes.

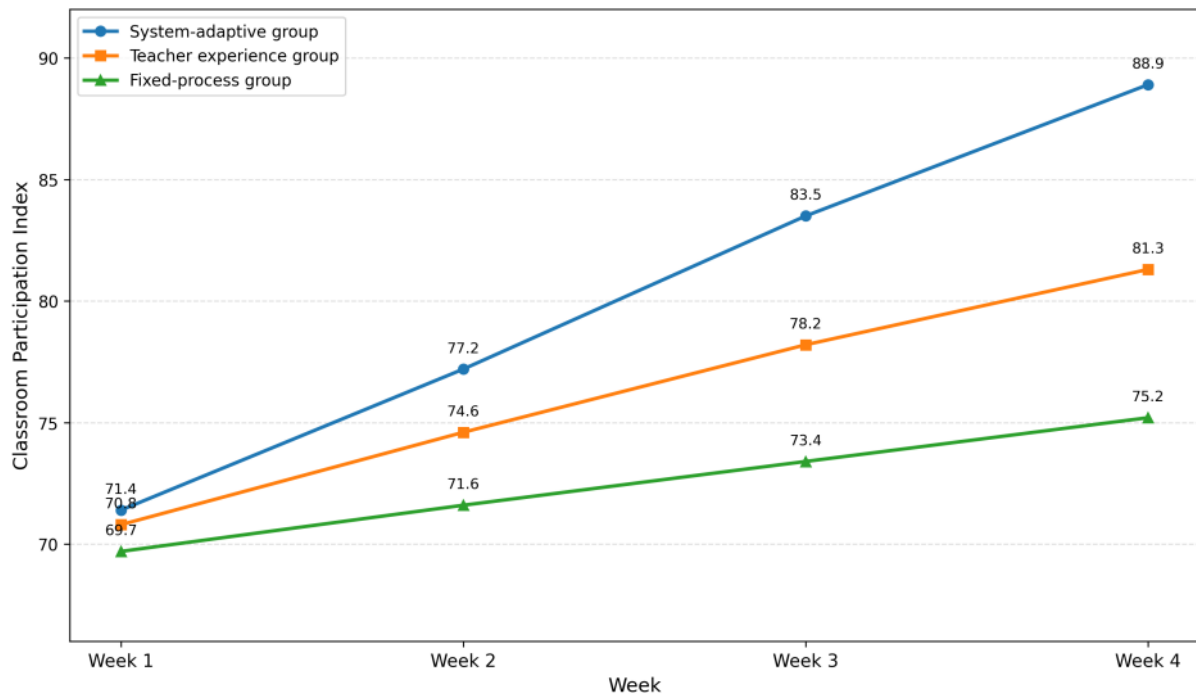


Figure 8: Weekly changes in the classroom participation index under different teaching methods

As shown in Figure 8, with the advancement of the experimental week, the system adaptation group showed the most obvious improvement in the classroom participation index, which increased from 71.4 in the first week to 88.9 in the fourth week, with an overall increase of 17.5. Teacher experience adjustment group increased from 70.8 to 81.3, an increase of 10.5; The fixed process teaching group only increased from 69.7 to 75.2, with a relatively limited change. The results show that the dynamic adjustment based on the results of emotional intelligent recognition can gradually enhance students 'investment level in melody modeling, lyrics understanding and coordination in continuous teaching, and this gain is not a one-time fluctuation, but has certain stability and accumulation.

Table 7: Statistical results of classroom teaching adaptation effect

Teaching Mode	Strategy Matching Rate / %	Average Response Latency / s	Post-class Task Completion Rate / %	Emotional Engagement Improvement / %	Teaching Satisfaction / 5
System Adaptation Group	91.8	1.84	89.5	18.7	4.58
Teacher Experience Adjustment Group	84.1	3.27	82.4	12.5	4.21
Fixed-process Teaching Group	72.6	4.95	74.8	6.9	3.84

From the comprehensive statistical results, the system adaptation group is better than the other two groups in a number of indicators, and the relevant results are shown in Table 7. The

strategy matching rate reached 91.8%, which was significantly higher than 84.1% of the teacher experience adjustment group and 72.6% of the fixed process teaching group. The average response delay is only 1.84 s, which indicates that the system can complete the state judgment and suggestion output in a short time. At the same time, the completion rate of after-class tasks in the system adaptation group reached 89.5%, teaching satisfaction was 4.58, and the improvement of emotional engagement was 18.7%, all showing good classroom adaptation effects. From the classroom records and strategy output results, when the system identified that students had delay in lyrics understanding or rhythm following, it could usually trigger lightweight interventions such as semantic supplement, demonstration replay or collaborative reinforcement quickly. In general, the classroom teaching adaptation mechanism constructed in this paper not only has high strategy generation accuracy, but also has good classroom enforceability and continuous intervention value.

4.4 Analysis of comparative experiments and ablation experiments

In order to further verify the comprehensive advantages of the proposed method in the intelligent emotion recognition task of Belarusian folk music classroom, this paper sets up multiple sets of comparison experiments under the same data division and training environment, and combines ablation experiments to investigate the contribution of each component module to the final results. The comparison objects include the traditional machine learning model SVM, random forest, the temporal deep model BiGRU, the single Transformer, and the multi-modal fusion model proposed in this paper. In order to enhance the robustness of the comparison experiment, a five-fold cross validation is used on the basis of the unified data in 4.4, and the results of the comparison experiment are shown in Figure 9(a). It can be seen that the performance of traditional models in complex classroom situations is relatively limited. The Accuracy of SVM is 79.46%, Macro-F1 is 77.92%, and the training time of a single round is 6.8s. The Accuracy of random forest is improved to 82.37%, and Macro-F1 is 80.84%, but the distinction between adjacent states is still not stable. After the introduction of time series modeling, the Accuracy of BiGRU reaches 86.21%, and Macro-F1 reaches 85.33%. Transformer is further improved to 88.37% and 87.64%, indicating that the attention mechanism has a good ability to capture the emotional state transfer in classroom clips. In contrast, the multimodal fusion model proposed in this paper achieves the best results in Accuracy, Macro-F1 and comprehensive robustness, with Accuracy reaching 91.62% and Macro-F1 reaching 90.84%, although the training time increases to 34.8 s/epoch. However, it is still within the acceptable range of classroom deployment.

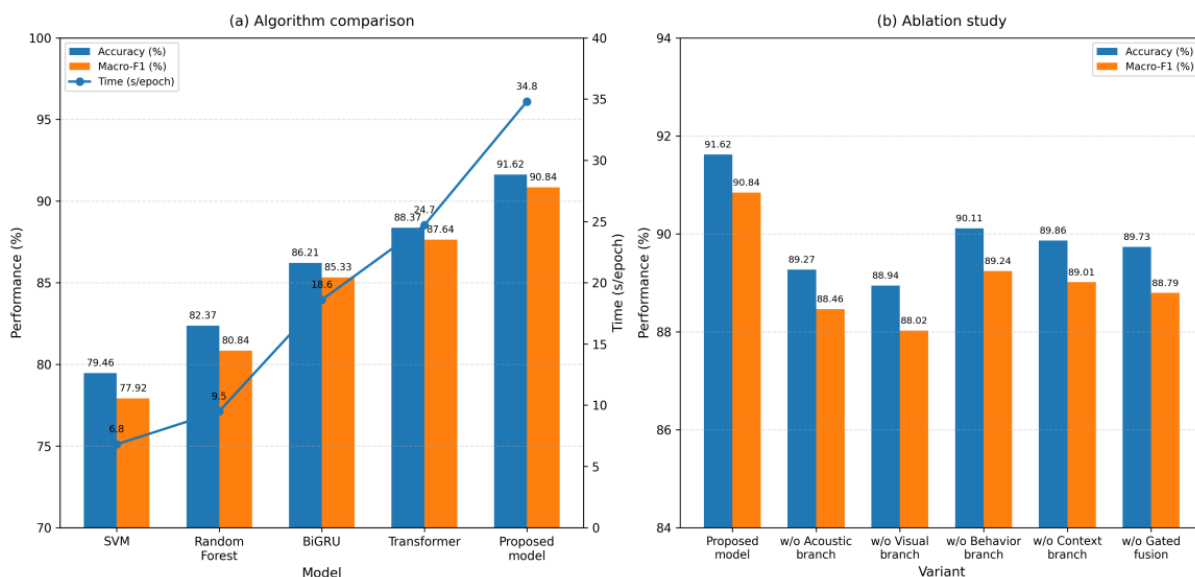


Figure 9: Comparison of different models and ablation experimental results

In order to further explain the source of model performance improvement, this paper continues to carry out ablation experiments, removing the acoustic channel, visual channel, behavioral channel, situational channel and gated fusion module respectively, and the results are shown in Figure 9(b). After removing the visual channel, the Accuracy of the model decreases to 88.94%, and Macro-F1 decreases to 88.02%, which shows that the information of students' facial expressions, gaze changes and posture has a high weight in the intelligent recognition of classroom emotions. After removing the acoustic channel, the Accuracy is reduced to 89.27%, and the Macro-F1 is 88.46%, indicating that the fluctuation of fundamental frequency, energy and rhythm in singing is still an important basis for judging the emotional engagement and regulation state. After removing the behavior channel and situation channel, the Accuracy decreased to 90.11% and 89.86%, respectively, indicating that the classroom interaction rhythm and the semantics of teaching fragments could provide necessary supplement for the state interpretation.

4.5 Discussion

Based on the above experimental results, it can be seen that the emotional intelligence state in the Belarusian folk music classroom does not originate from a single explicit signal, but is a composite result formed by the joint action of sound expression, expression posture, classroom behavior and teaching context. The key reason why the multimodal fusion model is superior to various single-channel methods is not only that the input information is more, but also that there is a complementary relationship between different modalities. Acoustic features can reflect the emotional tension and control level of students in the process of singing and responding, visual information is more sensitive to attention, collaboration willingness and immediate feedback, and behavior logs and situation labels provide necessary classroom background for state interpretation. Therefore, the effectiveness of the proposed method is essentially reflected in the better adaptation to the characteristics of "multi-source, dynamic and context dependent" in the folk music classroom. From the perspective of teaching, the improvement of strategy matching rate, response delay and classroom participation of the system adaptation mechanism shows that the intelligent emotion recognition is not only of analytical value, but also can be transformed into an executable classroom support tool. Especially in scenes such as lyrics understanding lag, rhythm follow-up imbalance, and weak

coordination, model-driven lightweight intervention is more timely than relying solely on empirical observation, and it is easier to form a continuous feedback chain. This shows that embedding computational models into folk music teaching does not mean that teachers' main role is weakened, on the contrary, it provides teachers with more fine-grained state reference and makes teaching adjustment more targeted. However, this paper still has some limitations. First, the data sources mainly focus on a limited number of classes and works. Although a variety of teaching links have been covered, the adaptability to the internal more subdivided styles of Belarusian folk music still needs to be verified. Secondly, although the multi-dimensional scoring mechanism is introduced into the emotional intelligent labeling, the real emotional activities in the classroom are fuzzy and fluid, and the existing annotations are still difficult to fully cover the complex psychological changes. Third, the system is more suitable for auxiliary feedback and cannot replace teachers' comprehensive judgment on cultural semantics and artistic expression. Subsequent research can be further deepened in the aspects of cross-school data expansion, weakly supervised labeling, long-term individual tracking and interpretable interactive interface, so as to improve the generalization ability and continuous application value of the system in real music education scenarios.

5 Conclusion and Prospect

Focusing on the problem of emotional intelligent recognition and teaching adaptation in Belarusian folk music classroom, this paper constructs a complete technical path composed of multi-source classroom data collection, multi-modal state modeling, strategy generation to real-time feedback. The study incorporated singing audio, facial expression, classroom behavior and teaching situation into the unified representation space, which alleviated the problem that traditional classroom evaluation relied on empirical observation and was difficult to continuously capture emotional changes. The experimental results show that the Accuracy of the proposed multi-modal fusion model reaches 91.62% and Macro-F1 reaches 90.84%, which are significantly better than that of the single model. In the classroom deployment test, the strategy matching rate of the system reaches 91.8%, the average response time delay is 1.84 s, and the teaching satisfaction score reaches 4.58, which shows that the method not only has good recognition accuracy, but also has the application feasibility in the real classroom. It should be pointed out that the existing data samples are still mainly from limited classes and works, and the adaptability of the model to more complex style differences and long-term individual changes still needs to be further tested. Subsequent research can be further deepened in the aspects of cross-school data expansion, weakly supervised labeling, lightweight deployment and interpretable interactive interfaces, so as to form a more stable intelligent collaboration mechanism between emotional intelligent recognition and folk music teaching.

References

- [1] Wang C, Zheng Z. Emotion recognition of Chinese traditional folk music using an assembling machine learning method[C]//Proceedings of the 2022 7th International Conference on Machine Learning Technologies. 2022: 30-35.
- [2] Vaizman T. Teaching musical instruments during COVID-19: teachers assess struggles, relations with students, and leveraging[J]. Music Education Research, 2022, 24(2): 152-165.

- [3] Uludag A K, Satir U K. Seeking alternatives in music education: The effects of mobile technologies on students' achievement in basic music theory[J]. *International Journal of Music Education*, 2025, 43(2): 172-188.
- [4] Campbell P S, Mellizo J M. Teaching music/teaching culture: From the rhetorical to the realities[J]. *Music Educators Journal*, 2024, 111(2): 26-34.
- [5] Nissen J. Aspirations and limitations: the state of world music education in secondary schools in multicultural Manchester[J]. *British Journal of Music Education*, 2023, 40(3): 385-396.
- [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] Han X, Chen F, Ban J. Music emotion recognition based on a neural network with an Inception-GRU residual structure[J]. *Electronics*, 2023, 12(4): 978.
- [8] Zhao J, Ru G, Yu Y, et al. Multimodal music emotion recognition with hierarchical cross-modal attention network[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022: 1-6.
- [9] Liang J. Developing emotional intelligence in a static and interactive music learning environment[J]. *Frontiers in Psychology*, 2024, 15: 1279530.
- [10] Váradi J, Szűcs T, Kerekes R, et al. A Systematic Review on the Emotional Dimensions of Music Education[J]. *Harmonia: Journal of Arts Research and Education*, 2024, 24(2): 236-246.
- [11] Culp M E, Svec C, McConkey M, et al. Meeting the social and emotional needs of P–12 learners: A descriptive study of music teacher education programs[J]. *Journal of Research in Music Education*, 2024, 72(1): 5-27.
- [12] Tu J, Fu H. The path to happiness for music students: Music empathy and music engagement as potential sources of subjective well-being[J]. *Humanities and Social Sciences Communications*, 2024, 11(1): 1-9.
- [13] Ouyang M. Employing mobile learning in music education[J]. *Education and Information Technologies*, 2023, 28(5): 5241-5257.
- [14] Maharaj A, Gill A. Technology in music education[J]. *Canadian Journal of Learning and Technology*, 2023, 49(2): 1-15.
- [15] Liu X, Shao X. Modern mobile learning technologies in online piano education: Online educational course design and impact on learning[J]. *Interactive Learning Environments*, 2024, 32(4): 1279-1290.
- [16] Paschalidou S. Technology-mediated hindustani dhrupad music education: an ethnographic contribution to the 4E cognition perspective[J]. *Education Sciences*, 2024, 14(2): 203.
- [17] Cheng L, Moir Z, Bell A P, et al. Digital musicianship in post-pandemic popular music

- education[J]. International journal of music education, 2024: 02557614241287558.
- [18] Zhou W, Guo K, Ying Y, et al. Chinese local music teaching materials: A review from 1934 to 2022[J]. Social sciences & Humanities open, 2024, 9: 100742.
- [19] Kakimova L S, Balagazova S T, Mukeyeva N E, et al. Pedagogical Technologies for the Emotional Intelligence of a Music Teacher[J]. World Journal on Educational Technology: Current Issues, 2022, 14(3): 817-824.
- [20] Kuldanov N, Balagazova S, Ibrayeva K, et al. The Use of Data-Based Emotion Recognition Systems for the Development of Emotional and Musical Creativity in Vocal Students[J]. Journal of Educational Computing Research, 2025, 63(5): 1122-1144.