ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Learning Mode and Innovative Design of Intelligent multimedia Platform in Vocal Music Performance Teaching

Zhuo Zhang[1,*]

[1] Huanghe Science and Technology University, Zhengzhou, Henan, 450000 China

SUMMARY: *With the continuous advancement of the digital transformation of vocal performance teaching, the shortcomings of traditional platforms in resource organization, learning support and feedback timeliness have become increasingly apparent. This paper constructs an intelligent multimedia platform, uses the multi-modal representation method of audio, video, text and interaction log fusion, combines learning behavior analysis to realize the generation of personalized learning paths, and designs a feedback closed-loop control mechanism. The experiment was carried out based on 96 students, 6 teachers and 3600 groups of effective samples. The results show that the average response time of the platform is 0.82 s, the recommendation accuracy is 91.6%, and the system availability is 99.1%. The comprehensive performance of the experimental group increased from 74.6 to 86.8, the learning engagement index increased from 0.61 to 0.83, and the repetition error rate decreased to 6.9%. The research shows that the platform can improve the precision support ability and training effect of vocal music teaching, and provide reference for the intelligent upgrading of vocal music performance teaching.*

KEYWORDS: *Vocal performance teaching; Multi-modal feature modeling; Intelligent recommendation; Optimization of Teaching Feedback*

1 Introduction

The teaching of vocal music performance has multiple attributes such as skill training, aesthetic perception and emotional expression. The teaching effect depends not only on the teacher's demonstration and explanation, but also on the practice process monitoring, the timeliness of learning feedback, and the degree of resource adaptation. Traditional classroom has long relied on face-to-face demonstration, repeated singing and self-practice after class, which is direct in the basic training stage, but there are also some practical limitations. On the one hand, it is difficult to continuously record the practice process of students after class, and teachers often can only make judgments based on the stage display results, so it is difficult to accurately grasp the change of vocal state, rhythm stability and emotional expression deviation. On the other hand, teaching resources mostly exist in scattered video, audio and text materials, and there is a lack of effective correlation between the contents. Students of different ability levels often have problems such as mismatched training intensity, unclear learning path and feedback lag when facing homogeneous training arrangements.

In recent years, multimedia technology, learning analysis technology and intelligent recommendation method continue to enter the education scene, which provides new conditions for the optimization of vocal performance teaching mode. Different from single resource

*zhangzhuo198511@163.com
https://doi.org/10.65102/is2026993
Published: 30 April, 2026

playing platform, intelligent multimedia platform emphasizes the collaborative organization of multi-source information such as audio, video, text and behavior log, which can record students 'practice frequency, work completion, error types and interaction tracks in the teaching process, and form more targeted learning support. For vocal music performance, singing demonstration, example explanation, action prompt, emotional interpretation and practice feedback themselves have distinct multimodal characteristics. Introducing computer technology into platform design not only helps to improve the efficiency of teaching resource management, but also helps to promote the learning mode from experience-driven to data-assisted driven.

Based on this, this paper focuses on the application of intelligent multimedia platform in vocal music performance teaching, and tries to construct a learning mode and innovative design scheme that takes into account resource representation, learning path recommendation and feedback regulation. The research focuses on the overall architecture of the platform, the modeling of multimodal teaching resources, the generation of personalized learning paths and the optimization of feedback closed-loop. The application value of the platform in teaching support and learning effect improvement is tested through experimental design, in order to provide new technical ideas and practical reference for the digital and intelligent development of vocal performance teaching.

2 Related Work

In recent years, research on intelligent teaching platforms, online learning environments, and multimedia resource organization has continued to increase. Zhang and Zhang (2024) focused on vocal teaching in the context of internet remote learning, proposing a new vocal teaching model based on online platforms, multimedia tools, and real-time interaction, and pointed out that this model has good performance in music theory learning, singing skill improvement, and learning satisfaction [1]. Qin (2024) discussed the design ideas of video teaching systems from the perspective of computer 5G technology and advanced algorithms, explaining that high-speed network transmission and algorithm support are crucial for the stable operation of video-based teaching platforms [2]. Kivuti et al. (2024) further from the perspective of learner experience quality, constructed an interactive multimedia model for streaming media content, and the research showed that its solution could improve page loading speed and video playback quality, providing a technical reference for the media distribution optimization of teaching platforms [3]. Wu et al. (2021) combined artificial intelligence with multimedia teaching platforms, discussed the integration path of university course integration in the platform, indicating that intelligent technology has moved from single resource display to collaborative teaching organization and process [4].

Based on this, the research perspective has expanded from platform construction to learning analysis, interaction experience, and intelligent recommendation. Wang (2025) introduced the hybrid data clustering algorithm into the simulation scenarios of e-learning courses, demonstrating that data mining methods have potential applications in the classification of learning behaviors and the optimization of interactive experiences [5]. Liu and Li (2025) discovered through a quasi-experimental study that the introduction of multimedia and interactive platforms helps improve students' grades and learning motivation, indicating that the value of the platform lies not only in the technical aspect but also in the teaching effect aspect [6]. Chen et al. (2024) analyzed the AI-assisted multimedia creation platform, pointing out that the abilities of speech recognition, visual perception, and intelligent behavior can enhance classroom participation and understanding levels [7]. Bajahzar (2024) started from the reform of higher education platforms and emphasized the supporting role of AI models in the

optimization of multimedia education systems [8]. Zhang (2024) designed a platform based on deep learning in English-Chinese translation teaching and improved the intelligence level of teaching interaction through feature embedding and classification methods [9]. Yan and Liu (2023) designed a university English intelligent teaching platform based on large-scale multimedia data technology, emphasizing the construction of an autonomous learning environment centered on learners [10]. Overall, the existing research has provided basic support such as platform architecture, media transmission, interaction design, and algorithm support for the construction of intelligent teaching platforms, but for this strong practice, strong feedback, and strong multimodal coupling scenario of vocal performance teaching, there is still a lack of systematic research that unifies the representation of audio, video, text resources, learning path recommendations, and the closed-loop of teaching feedback, which also constitutes the entry point for further research in this paper.

3 Intelligent Multimedia Platform Learning Mode Construction and Innovative Design Method for Vocal Performance Teaching

3.1 Overall Architecture and Function Module Design of the Vocal Performance Teaching Multimedia Platform

To enhance the synergy of resource invocation, learning support, and feedback regulation in vocal performance teaching, this paper constructs an intelligent multimedia platform overall architecture for the entire teaching process. The overall architecture of the vocal performance teaching intelligent multimedia platform is shown in Figure 1.

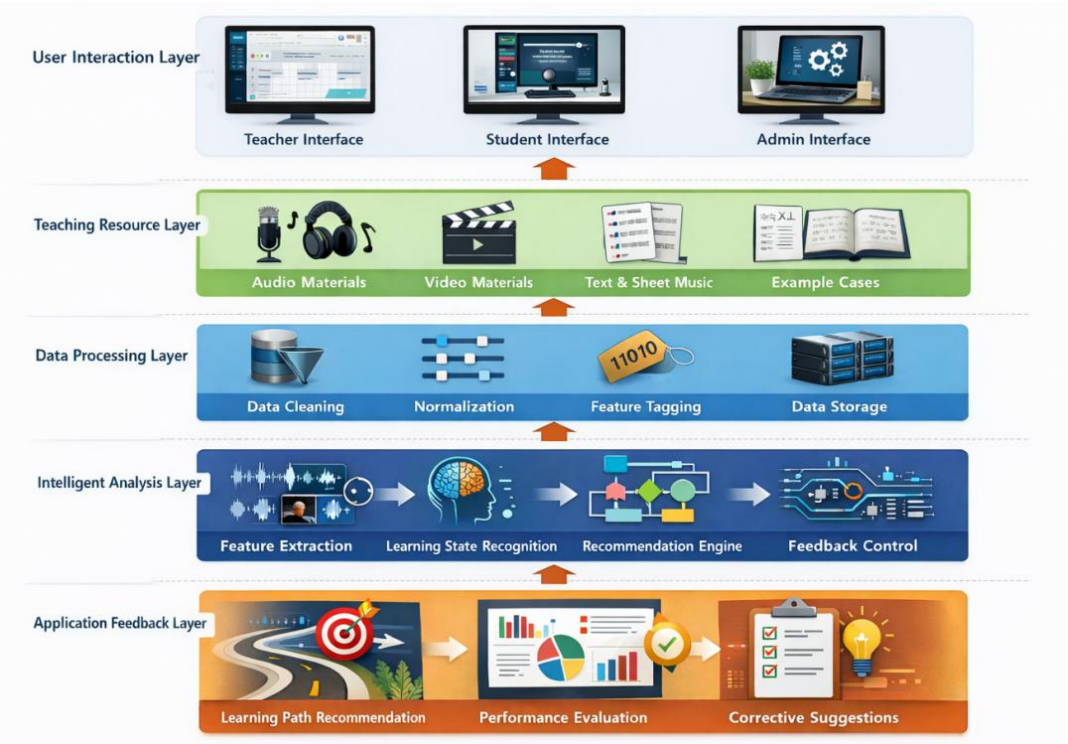

Figure 1: Overall Architecture Diagram of the Intelligent Multimedia Platform for Vocal Music Teaching

The platform is not a simple video playback or data storage system, but forms a closed-loop operation mechanism around "resource organization-behavior perception-intelligent analysis-teaching feedback", so that the teaching design of the teacher side, the learning activities of the student side and the algorithm support of the platform side can be completed in a unified environment. Combined with the requirements of demonstration, repeatability and instant correction in vocal music teaching, the platform is divided into five parts: user interaction layer, teaching resource layer, data processing layer, intelligent analysis layer and application feedback layer. The user interaction layer mainly undertakes the functions of teachers publishing tasks, students submitting exercises, online evaluation and learning record display. The teaching resource layer is responsible for integrating audio, video, text, score and case materials. The data processing layer completed multi-source data cleaning, standardization and label mapping. The intelligent analysis layer focuses on feature extraction, learning state recognition and recommendation decision-making. The application feedback layer outputs learning paths, training suggestions and effect evaluation results for teachers and students.

In order to describe the collaborative relationship between the internal modules of the platform, the comprehensive running state of the multimedia platform for vocal music teaching at time t is set as follows.

$$P_t = \{U_t, R_t, D_t, A_t, F_t\} \tag{1}$$

Among them, U_t represents the user interaction status, Rt represents the teaching resource organization status, D_t represents the data processing status, A_t represents the intelligent analysis status, and Ft represents the feedback output status. To measure the support intensity of the platform for a single learner, the platform service efficiency function is defined as:

$$E_i = \alpha S_i + \beta M_i + \gamma Q_i \tag{2}$$

where, E_i represents the comprehensive support effectiveness of the platform for learner i, Si represents the resource matching degree, M_i represents the interactive response efficiency, Q_i represents the feedback quality, α, β, γ are weight coefficients, and satisfies:

$$\alpha + \beta + \gamma = 1 \tag{3}$$

The formula can be used to describe the overall performance of the platform in three dimensions of resource supply, process response and feedback output. When the matching degree between learners 'exercise content and resource recommendation is high, the interaction delay is low, and the feedback results are more accurate, the service efficiency of the platform will be improved.

In terms of function implementation, the teacher end focuses on teaching arrangement, task configuration, resource management and learning monitoring, the student end focuses on demonstration viewing, segmented practice, assignment uploading and result viewing, and the algorithm end of the platform continues to perform fusion analysis of practice audio, singing video, text evaluation and behavior log. Therefore, the platform is no longer limited to static resource display, but turns to an intelligent teaching environment that can perceive the learning process, identify different needs and generate dynamic support, which provides a unified structural foundation for subsequent multi-modal feature modeling and personalized learning path recommendation.

3.2 Multi-modal Teaching Resource Representation and Feature Modeling Incorporating Audio, Video, and Text

Musical teaching resources are not a simple stacking of a single medium. A complete singing learning process often includes multiple types of information such as sound fluctuations, mouth shape and body posture changes, song meaning and sheet music prompts. If audio, video, and text are still stored and retrieved separately in the traditional way, the platform can only complete resource display but is unable to further identify the subtle deviations in the learner's practice. Based on this, this paper regards musical teaching resources as multi-modal objects with temporal correlation and semantic correspondence, and completes representation and modeling under a unified computing framework, providing computable input for subsequent learning state recognition and personalized recommendations.

In terms of audio modalities, the fundamental frequency, short-term energy, formant distribution, Mel-frequency cepstral coefficient and rhythm stability are extracted to reflect pitch control, breath support and vocal coherence. Let the audio feature vector of the t-th time slice be:

$$a_t = [f_t, e_t, c_t, r_t] \tag{4}$$

Among them, f_t represents the fundamental frequency feature, e_t represents the energy intensity, ct represents the spectral cepstrum feature, and rt represents the rhythm deviation quantity. This vector can more comprehensively describe the acoustic performance of the student during the singing process.

In the video modalities, this paper does not simply capture the singing image, but combines the mouth shape area, head posture, chest and abdomen movement, and body movements to construct dynamic visual features. Let the video modalities at time t be represented as:

$$v_t = \psi(m_t, p_t, g_t) \tag{5}$$

where, m_t is the mouth shape opening and closing feature, pt is the posture information, gt is the expression and action change feature,

$\psi(\cdot)$ represents the visual encoding function. The significance of this processing lies in that the platform can not only see "what was sung", but also judge "how it was sung", especially suitable for discovering situations where the vocalization action does not match the audio result.

The text modalities mainly correspond to the lyrics content, example explanations, teacher annotations, and training prompts. Considering that the text information has a strong semantic constraint effect, this paper uses the context encoding method to generate text vectors:

$$t_t = \phi(w_1, w_2, \ldots, w_n) \tag{6}$$

Among them, w_1, w_2, ..., w_n represent the input word sequence, and $\phi(\cdot)$ represents the text semantic mapping function. The introduction of text vectors enables the platform to convert teaching prompts such as "emotional processing", "breathing position", and "articulation requirements", which were originally difficult to quantify, into semantic features that can be involved in calculation.

To reduce the interference caused by inconsistent dimensions and temporal misalignment between different modalities, this paper further constructs a gated fusion mechanism, mapping audio, video, and text features to the same latent space. The fusion expression is as follows:

$$h_t = \sigma(W_g[a_t; v_t; t_t] + b_g) \odot (W_a a_t + W_v v_t + W_t t_t) \quad (7)$$

where $\sigma(\cdot)$ is the activation function, $[\cdot]$ represents vector concatenation, $\odot$ represents element-wise multiplication, and ht is the unified multimodal representation. Compared with direct stitching, the proposed method is able to adaptively adjust the weights according to the feature contributions of different teaching segments. For example, the weight of audio modality is increased in the breath training segment, the influence of video modality is enhanced in the stage performance training segment, and the role of text modality is highlighted in the stage of work understanding.

To ensure the consistency of cross-modal representation, alignment constraints are introduced as follows.

$$L_{align} = \sum_{t=1}^{T} (\|W_a a_t - W_v v_t\|_2^2 + \|W_v v_t - W_t t_t\|_2^2) \quad (8)$$

This formula is used to compress the semantic offset of different modalities at the same teaching moment, so that the unified representation not only retains their own information, but also does not split each other. After the above processing, the platform can form a multi-modal resource feature matrix for vocal music teaching scenarios, which not only supports resource retrieval and content matching, but also provides a more fine-grained data basis for subsequent learning path generation, performance evaluation and feedback regulation. From the perspective of teaching application, this modeling method makes "auditory performance, visual action and text instruction" be jointly expressed in the same computing link for the first time, which better reflects the technical characteristics and innovative design ideas of the intelligent platform for vocal music performance teaching.

3.3 Personalized Learning Path Generation and Recommendation Method Based on Learning Behavior Analysis

In the teaching scenario of vocal performance, the differences between students are not only reflected in the range condition and basic level, but also in the behavior level of practice frequency, error type, resource preference, feedback response speed, and stage progress. If the platform still adopts the unified task push method, it is easy to have the problem that students with weak foundation cannot keep up and students with strong ability repeat too much training. To this end, this paper builds the generation of personalized learning path on the basis of learning behavior analysis. By calculating the logs of students 'clicks, stays, exercise submissions, repeated plays, error correction times and result fluctuations in the platform, the dynamic learning portraits are formed, and the learning units are sorted and recommended.

Let the behavior sequence of learner i within period T be:

$$B_i = \{b_1, b_2, \cdots, b_T\} \quad (9)$$

Among them, "b_T" represents the record item of the t-th learning activity, which includes indicators such as resource access duration, practice completion rate, error count, re-practice frequency, and feedback response. Based on the behavior sequence, a learner state vector is constructed:

$$z_i = [f_i, e_i, p_i, c_i, s_i] \quad (10)$$

Among them, f_i represents the practice frequency, e_i represents the error density, p_i represents the progress rate, c_i represents the content preference, and s_i represents the stability coefficient. This vector can reflect the current learning status and trend of the student relatively compactly.

To avoid the recommendation results remaining at the level of "similarity resource push", this paper constructs the vocal training content as a directed learning graph G=(V,E). Here, the nodes V represent learning units such as vocalization training, breath control, rhythm imitation, work expression, and stage presentation, and the edges E represent the prerequisite and transfer relationships between each unit. For any candidate learning unit vj, the adaptation score of it to the learner i is defined as:

$$\begin{aligned} \mathrm{Score}(i,j) = \alpha \cdot \mathrm{Sim}(z_i, r_j) + \beta \cdot \mathrm{Gap}_{ij} + \\ \gamma \cdot \mathrm{Gain}_j - \delta \cdot \mathrm{Cost}_j \end{aligned} \tag{11}$$

where, rj is the feature representation of learning unit j, Sim(·) represents the matching degree between learner state and resource characteristics, Gapij represents the compensation degree of the unit for the current ability shortboard, Gainj represents the expected learning benefit, Costj represents the time required to complete the unit and cognitive load, α,β,γ,δ are weight coefficients. This formula takes into account the three levels of "whether it is suitable for learning", "whether it is worth learning" and "whether it will learn too much", so that the recommendation process is more in line with the real teaching needs.

On this basis, the platform aims to maximize the cumulative revenue and generates the optimal path of the learner:

$$P_i^* = \arg\max_{P_i} \sum_{j=1}^{L} S\,\mathrm{core}(i,j) - \lambda \sum_{j=2}^{L} |\, d_j - \hat{d}_i| \tag{12}$$

where P_i^* represents the optimal learning path of learner i, L is the path length, dj is the difficulty of the JTH unit, $\hat{d}_i$ is the current tolerable difficulty level of learner, and λ is the difficulty penalty coefficient. Instead of simply pushing all the high-yield content to the student, the platform controls the training gradient to keep the learning path consistent and executable.

To enhance the dynamic adaptability of the recommendation results, this paper further introduces a rolling update mechanism. When a student completes a learning unit, the platform re-estimates the state vector zi based on the latest submitted audio, video, and test results, and then calculates the scores of subsequent nodes to achieve "learning while adjusting" path correction. Compared with static course scheduling, this mechanism is more suitable for the training characteristics of vocal performance teaching with obvious phased fluctuations and high feedback dependence.

Table 1: Learninger Behavior Characteristics and Personalized Path Recommendation Strategy

Learninger Type	Primary Behavioral Characteristics	Recommended Learning Content	Path Intensity	Expected Goal
Weak Foundation Type	Low practice frequency, high error density, many repetitions	Accurate pitch imitation, basic vocalization, phrase singing	Low to Medium	Establish basic vocal stability
Rhythm Imbalance Type	Large rhythm deviation, many submissions but slow improvement	Beat training, segmented rhythm correction, accompaniment practice	Medium	Reduce rhythm fluctuation
Expression Deficiency Type	Stable pitch, but low emotional score	Work understanding, song meaning training, expression action demonstration	Medium	Improve emotional expression completeness
Skill Enhancement Type	High completion rate, fast progress rate, active feedback response	High difficulty works, stage presentation, comprehensive singing tasks	Medium to High	Expand expressiveness and transfer ability
Fluctuating and Repeated Type	High stage performance fluctuations, low stability coefficient	Error analysis, short-term consolidation, targeted error correction training	Medium	Improve performance consistency

Overall, this section's method is not simply based on historical clicks for recommendation, but integrates learning behavior, ability shortcomings, resource benefits, and training load into the calculation process, making the generation of personalized learning paths more interpretable and adjustable. This method can not only perform intelligent scheduling on the service side, but also provide teachers with intuitive teaching stratification basis, thereby laying the foundation for subsequent teaching feedback optimization and intelligent control mechanism design.

3.4 Platform Interaction Optimization and Intelligent Control Mechanism Design for Teaching Feedback Cycles

Vocal performance teaching is not a one-way process of "resource push-student finish-teacher evaluation". What really affects the quality of learning is often whether the deviation can be found in time after practice, the nature of the deviation can be judged, and the subsequent tasks and interaction methods can be adjusted accordingly. Based on this understanding, this paper introduces an intelligent regulation mechanism for teaching feedback closed loop on the platform side, which integrates students 'exercise results, system recognition results, teachers 'evaluation opinions and platform response records into the same regulation link, so that the platform can continuously revise the interaction strategy according to the dynamic changes in the learning process, rather than waiting for the static summary after the end of the unit.

Let the integrated feedback bias of learner i after round t of practice be as follows.

$$R_i^{(t)} = \lambda_1 \left| \hat{y}_i^{(t)} - y_i^{(t)} \right| + \lambda_2 \Delta r_i^{(t)} + \lambda_3 \left(1 - c_i^{(t)}\right) \tag{13}$$

Here, $y_i^{(t)}$ represents the predicted performance value of the platform, $y_i^{(t)}$ represents the actual measured performance value, $\Delta r_i^{(t)}$ represents the feedback delay, $c_i^{(t)}$ represents the credibility of the current evaluation, and λ1,λ2, and λ3 are the adjustment coefficients. This formula not only focuses on "whether the song is accurate or not", but also takes into account whether the feedback is timely and whether the judgment is reliable, so as to avoid excessive intervention of the platform on the basis of a single score.

In order to distinguish general fluctuations from abnormal states that require immediate intervention, this paper further defines the dynamic trigger threshold:

$$\Theta_i^t = \Theta_0 + \mu \sigma_i^t - \nu g_i^{(t)} \tag{14}$$

Here, Θ0 is the base threshold, σ_i^t represents the degree of recent performance volatility, $g_i^{(t)}$ represents the extent of stage progress, μ and ν are the weights. The advantage of this design is that the platform will not interrupt students frequently because of a minor mistake, nor will it ignore the problems that continue to accumulate, and it can strike a balance between "no overintervention" and "timely correction".

After triggering the control, the platform does not use a fixed template-style feedback, but generates differentiated interaction actions based on the source of the deviation and the learning state. Let the next round of control vector be:

$$u_i^{(t+1)} = \tanh\left(W_r R_i^{(t)} + W_s s_i^{(t)} + W_m m_i^{(t)} + b\right) \tag{15}$$

Among them, $s_i^{(t)}$ represents the learning state vector, $m_i^{(t)}$ represents the current task modal feature, and $u_i^{(t+1)}$ outputs the set of interactive control actions, including resource rearrangement, difficulty rollback, segmented demonstration enhancement, key error correction prompts, and teacher review requests, etc. Since the causes of different problems in vocal music training are not the same, the feedback methods corresponding to pitch deviation, unstable breath, insufficient emotional expression, and imbalance in action control should also be different. Therefore, this control vector is essentially an adaptive decision expression oriented towards the teaching scenario.

To keep the closed-loop mechanism working continuously, the platform also updates the resource priorities on a rolling basis. For a candidate resource k, its scheduling weight in the next round is defined as follows.

$$\pi_k^{(t+1)} = \frac{\exp\left(z_k^{(t)} + \rho a_{ik}^{(t)} - \xi l_k\right)}{\sum_{j=1}^{K} \exp\left(z_j^{(t)} + \rho a_{ij}^{(t)} - \xi l_j\right)} \tag{16}$$

Here, $z_k^{(t)}$ is the resource base score, $a_{ik}^{(t)}$ represents the adaptation of the resource to the current learner problem, lk represents the resource load cost, and ρ and ξ are control parameters. This formula means that when the platform generates the subsequent training content, it no longer looks at the popularity of the resource itself, but pays more attention to its ability to repair the current problem and the cost of use.

The closed-loop process thus formed can be summarized as: the platform first receives the

practice results, then judges the degree of deviation and credibility, subsequently decides whether to trigger control, and dynamically generates interactive actions and resource rearrangement plans, and finally continues to correct based on the results of the next round of learning. Compared with traditional platforms, the mechanism designed in this paper connects "identification - judgment - intervention - re-evaluation" into a continuous chain, shifting the teaching feedback from post-class summary to the learning process itself, and transforming the intelligent platform from a resource container to a teaching support system with diagnostic and control capabilities. This design is more in line with the characteristics of immediate performance in vocal music performance training, fast error transmission, and obvious individual differences, and provides a clear mechanism basis for the performance verification and teaching effect analysis of the platform in subsequent experiments.

4 Experimental Design

4.1 Construction and Sample Collection of the Multimodal Dataset for Vocal Music Performance Teaching

In order to verify the applicability of the built platform in the teaching scenario of vocal performance, this paper constructs a multimodal dataset for real teaching process. A total of 96 students and 6 teachers were included in the sample collection from music courses in two universities. The data of classroom teaching, after-class practice and stage test were collected for 12 weeks. The collection content covers singing audio, practice video, lyrics and sample text, teacher's annotations, and platform interaction logs, so as to restore the real behavior track of students in the process of "listening to demonstration - practice - error correction - practice again". To ensure data comparability, all audio is saved at a uniform sampling rate, video is recorded at a fixed resolution and frame rate, and textual material is structurally annotated by work, training task, and feedback type. After the collection is completed, the samples with missing frames, distortion, timing misalignment and incomplete annotation are removed, and then the "exercise clip - feedback record - learning result" is used as the basic alignment unit to form a multi-modal sample set for subsequent training and testing. Considering the influence of different training contents on model learning, five tasks including basic vocalization, breath control, rhythm imitation, song singing and comprehensive performance were retained in the data set, which made the sample distribution closer to the actual teaching situation. The composition and collection of multimodal data sets for vocal performance teaching are shown in Table 2. Finally, a total of 3600 groups of aligned effective multimodal samples were obtained, of which the training set, validation set and test set accounted for 70%, 15% and 15%, respectively, which can meet the needs of subsequent feature modeling, path recommendation and feedback control experiments.

Table 2: Composition and Collection of the Vocal

Music Performance Multimodal Dataset Data category	Original sample size	Valid sample size	Main collection content	Source
Singing audio	4120	3864	Fundamental frequency, loudness, rhythm, and timbre changes	Classroom singing and after-class practice recordings
Practice video	4120	3780	Lip movement, posture, expression, and movement trajectory	Classroom videos and platform-uploaded videos
Text resources	2380	2268	Lyrics, example explanations, and teacher comments	Teaching resource library and annotation records
Interaction log	28640	27912	Clicks, stays, re-practice, review, submission	Automatically recorded by the platform
Aligned multimodal samples	3860	3600	Audio-video-text-log combined samples	Generated by unified timestamp alignment

4.2 Preprocessing and feature fusion methods for multimodal data

In order to reduce the noise interference in the original teaching data and improve the comparability between different modalities, this paper uniformly preprocesses the audio, video, text and interaction log data before model training. In the audio part, silence segment clipping, background noise suppression and amplitude normalization are completed, and then the audio part is segmented into short segments according to a fixed window to retain the information of pitch, rhythm and energy variation. In the video part, key frames are extracted, invalid frames are eliminated, portrait region is located and pose sequence is smoothed to avoid errors caused by jitter, occlusion and illumination change. The text data is processed by word segmentation, stop word filtering, term standardization, and semantic encoding, so that lyrics descriptions, teacher comments, and training tips can be converted into computable features. The interaction log is rearranged according to the timestamp, and the behavioral indicators such as click frequency, stay time, playback number and task completion are extracted. In order to eliminate the differences between different feature dimensions, this paper uses standardization processing:

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{17}$$

where,x_{ij} represents the original feature value of the i-th sample in the j-th dimension, μ_j and σ_j are the mean and standard deviation of the feature in that dimension, respectively.

Considering the differences in sampling frequency and time length of multimodal data, this paper aligns the sequences based on a unified time axis. For feature sequences of different lengths, a combination of linear interpolation and truncation completion is used for synchronous mapping. Subsequently, to highlight the contribution of effective modalities in different training tasks, an adaptive weighted fusion strategy is introduced:

$$F = \omega_a A + \omega_v V + \omega_t T + \omega_l L \tag{18}$$

where, A, V, T, and L represent the feature matrices of audio, video, text, and log, respectively,

ωa, ωv, ωt, and ωl are the corresponding weights, and they satisfy:

$$\omega_a + \omega_v + \omega_t + \omega_l = 1 \tag{19}$$

After the above processing, the original discrete data is transformed into unified structure, temporally aligned, and semantically relatively complete fused features, providing stable input for subsequent platform performance evaluation and recommendation experiments.

4.3 Design of Evaluation Indicators for Platform Performance and Teaching Effectiveness

In order to more comprehensively test the practical application value of the built platform, this paper sets up evaluation indicators from two dimensions of platform operating performance and teaching implementation effect. The former mainly investigates the responsiveness and stability of the system in the process of multi-modal data access, resource scheduling and intelligent recommendation, while the latter focuses on the changes of students 'training quality, learning engagement and stage performance. Such indicator Settings can reflect whether the platform "works", but also whether the platform "really helps teaching".

In terms of platform performance, the average response latency, recommendation accuracy rate, and system availability rate are selected as the core indicators. The average response latency is defined as:

$$T_{avg} = \frac{1}{N}\sum_{i=1}^{N}(t_i^{out} - t_i^{in}) \tag{20}$$

Among them, "t_i^{in}" and "t_i^{out}" respectively represent the entry time and completion time of the i-th request, and N represents the total number of requests. The recommendation accuracy rate is used to measure the degree of consistency between the content pushed by the platform and the actual learning needs. Its expression is:

$$P@K = \frac{1}{N}\sum_{i=1}^{N}\frac{|R_i^K \cap G_i|}{K} \tag{21}$$

Among them, R_i^K represents the top K learning units recommended by the platform for learner i, and G_i represents the actual effective learning set for him/her. The system availability is defined as:

$$A_{sys} = \frac{N_{ok}}{N_{all}} \times 100\% \tag{22}$$

Among them, N_{ok} represents the number of tasks that were completed successfully, and N_{all} represents the total number of tasks.

In terms of teaching effectiveness, this article measures it using the rate of score improvement and the learning engagement index. The rate of score improvement indicates the magnitude of the change in students' test scores during the stage:

$$I_{score} = \frac{\bar{S}_{post} - \bar{S}_{pre}}{\bar{S}_{pre}} \times 100\% \tag{23}$$

Among them, $\bar{S}_{pre}$ and $\bar{S}_{post}$ represent the average scores before and after the experiment respectively. The learning engagement index is constructed by integrating the click frequency, practice duration, and task completion rate:

$$E_{learn} = \frac{1}{N}\sum_{i=1}^{N}(\theta_1 c_i + \theta_2 d_i + \theta_3 q_i) \tag{24}$$

Here, c_i represents the click frequency of the learner, d_i represents the effective practice duration, q_i represents the task completion rate, and θ1, θ2, θ3 are weight coefficients. Through these indicators, the platform can be jointly evaluated from both technical performance and teaching benefits, providing quantitative basis for subsequent comparative experiments and result analysis.

4.4 Comparison Experiment Plan and System Implementation Environment Setup

In order to test the effectiveness of the method in the real teaching platform, the experiment is carried out by using the control method of unified data sources, unified training rounds and unified evaluation indicators. The comparison objects include not only the traditional teaching platform scheme, but also the improved method with certain intelligent recommendation ability. In the specific Settings, the experiment was divided into five groups. The first group was a traditional multimedia teaching platform, which only provided resources browsing, homework uploading and results viewing functions. The second group was a resource push method based on artificial rules, which distributed learning content to students according to the preset order of teachers. The third group is the recommendation method based on collaborative filtering, which generates learning unit recommendations according to similar learners' historical behaviors. The fourth group is an improved platform that only performs multi-modal feature fusion but does not have the ability of feedback closed-loop control. The fifth group is the personalized learning path generation and intelligent control platform proposed in this paper. Through this progressive comparison, the gains brought by multimodal modeling, path recommendation and feedback regulation on platform performance and teaching effect can be observed more clearly.

All experiments are based on the 3600 valid multimodal samples constructed in the previous text. The division of training set, validation set, and test set remains consistent, with proportions of 70%, 15%, and 15% respectively. To reduce the interference of accidental factors, each group of models uses the same batch size, learning rate, and maximum number of iterative rounds. During the training process, only differences in method structure and regulation mechanism exist. In terms of system implementation, the platform adopts a browser/server architecture, with the front end responsible for course display, task submission, practice record, and feedback viewing, and the back end responsible for data management, feature processing, recommendation scheduling, and result storage. The experimental running environment is configured as Ubuntu 22.04 operating system, Python 3.10 development environment, PyTorch 2.1 deep learning framework, and MySQL 8.0 for data storage and retrieval. The hardware end uses Intel Xeon processors, 64 GB memory, and NVIDIA RTX 4090 graphics cards to ensure the stability of the training and inference process.

During the experiment, students complete tasks such as vocal training, rhythm imitation, work performance, and comprehensive performance based on the learning path generated by the platform. Teachers only intervene and judge when there are anomalies in the system or the platform issues a review request, in order to as closely as possible simulate the platform support process in the real teaching scenario. Considering that the results in Chapter 5 will mainly

present the overall performance, recommendation effect, teaching effectiveness, and feedback regulation effect of the platform, this paper selects the first group of traditional multimedia teaching platform as the control group, and the fifth group of the personalized learning path generation and intelligent regulation platform proposed in this paper as the experimental group to observe the comprehensive gains of the intelligent platform in actual teaching applications; the second group to the fourth group are mainly used as progressive comparison objects for platform performance and recommendation effect analysis. This experimental arrangement retains the horizontal comparability between different methods and ensures the consistency of object setting for subsequent teaching effect analysis and feedback loop analysis, thus forming a clearer correspondence between the experimental design and result presentation.

5 Results

To verify the application effect of the proposed platform in vocal performance teaching, this paper conducts an analysis from four aspects: overall performance of the platform, personalized recommendation performance, changes in teaching effectiveness, and feedback loop control function.

(1) Overall performance comparison results of different platforms

The overall performance comparison results of different platform schemes are shown in Table 3. It can be seen that the method proposed in this paper outperforms the other comparison methods in terms of average response latency, recommendation accuracy, and system availability rate. Specifically, the average response latency of the proposed platform is reduced to 0.82 seconds, which is 44.2% shorter than that of the traditional multimedia platform (1.47 seconds); the recommendation accuracy reaches 91.6%, which is 14.8 percentage points higher than the manual rule-based push method; the system availability rate reaches 99.1%, indicating that the constructed platform maintains high stability under the conditions of multimodal data access and learning task scheduling. Overall, the platform architecture optimization and intelligent control mechanism jointly improve the system operation efficiency, providing reliable technical support for subsequent teaching applications.

Table 3: Overall performance comparison results of different platform schemes

Method	Average response latency/s	Recommendation accuracy/%	System availability rate/%
Traditional multimedia platform	1.47	72.4	95.8
Manual rule-based push platform	1.26	76.8	96.6
Collaborative filtering recommendation platform	1.05	84.7	97.8
Multimodal fusion platform	0.93	88.9	98.4
This method	0.82	91.6	99.1

(2) Analysis of personalized learning path recommendation effect

The performance trends of different recommendation methods are shown in FIG. 2. As the number of training rounds increases, Precision@10 of all methods shows an upward trend, but the proposed method converges faster and has smaller fluctuations in the later stage. At the 20th round of training, Precision@10 of the proposed method has reached 0.873, while that of the collaborative filtering method is only 0.798. At the 50th round, the proposed method is further improved to 0.916, which is 0.079 higher than the collaborative filtering method. This shows

that the path generation mechanism based on learning behavior analysis can more effectively identify students' current needs and reduce invalid resource push. Overall, the proposed method not only improves the recommendation accuracy, but also enhances the stability of the recommendation results in the continuous training process.

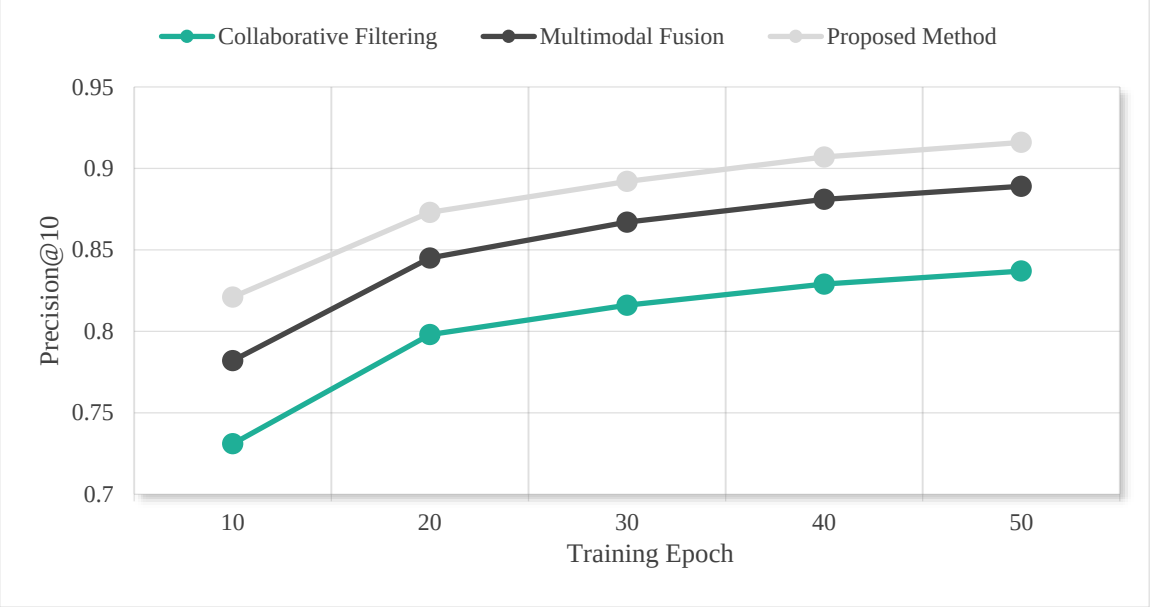

Figure 2: Line chart of Precision@10 changes for different recommendation methods

(3) Analysis of the improvement in teaching effectiveness

The results of the change in teaching effectiveness before and after the experiment are shown in Table 4. To ensure that the comparison objects for the teaching effectiveness are consistent with the experimental design in the previous section, this paper takes the students corresponding to the first group of traditional multimedia teaching platforms as the control group, and the students corresponding to the fifth group of the platform proposed in this paper as the experimental group. The changes in aspects such as pitch control, rhythm stability, emotional expression, comprehensive performance, and learning engagement were compared between the two groups within the same teaching cycle. After 12 weeks of training, the students in the experimental group showed more significant improvements in the four singing performance indicators. Among them, the comprehensive performance score increased from 74.6 points to 86.8 points, an increase of 12.2 points, calculated at the performance improvement rate defined in Section 4.3, it was approximately 16.4%; the learning engagement index increased from 0.61 to 0.83. In contrast, the comprehensive performance of the control group increased from 73.4 points to 79.6 points, an increase of 6.2 points, with a performance improvement rate of approximately 8.4%, and the learning engagement index only increased from 0.58 to 0.67. Thus, it can be seen that this platform not only improves the resource matching and feedback efficiency at the technical level, but also more effectively promotes the improvement of students' singing ability and learning participation at the teaching implementation level.

Table 4: Results of the change in teaching effectiveness before and after the experiment

Indicator	Control group before experiment	Control group after experiment	Experimental group before experiment	Experimental group after experiment
Pitch control/points	72.8	78.5	73.1	85.2
Rhythm stability/points	71.6	77.4	72.0	84.7
Emotional expression/points	70.9	76.1	71.3	85.6
Comprehensive performance/points	73.4	79.6	74.6	86.8
Learning engagement index	0.58	0.67	0.61	0.83

(4) Analysis of the effect of the feedback closed-loop regulation mechanism

The changes in the support effect of the platform before and after the feedback closed-loop regulation are shown in Figure 3. As the training weeks progress, the repetition error rate of students under the method proposed in this paper shows a continuous downward trend, while the rate of the platform without closed-loop regulation decreases relatively slowly. In the first week, the repetition error rates of the two groups were 18.7% and 18.5% respectively, and the difference was not significant; by the 12th week, the repetition error rate of the method proposed in this paper decreased to 6.9%, while the platform without closed-loop regulation remained at 10.8%, and the difference expanded to 3.9 percentage points. This result indicates that by integrating the identification of practice results, deviation judgment, resource rearrangement, and subsequent task adjustment into a unified feedback loop, the platform can more promptly capture the persistent problems of students in the training process and reduce the accumulation of repetitive errors through dynamic regulation. Combined with the platform operation logs, it can also be observed that after the implementation of the closed-loop regulation, the review pressure of teachers for repetitive problems has been reduced, indicating that this mechanism not only improves the efficiency of problem correction but also enhances the platform's proactive support ability in the teaching process.

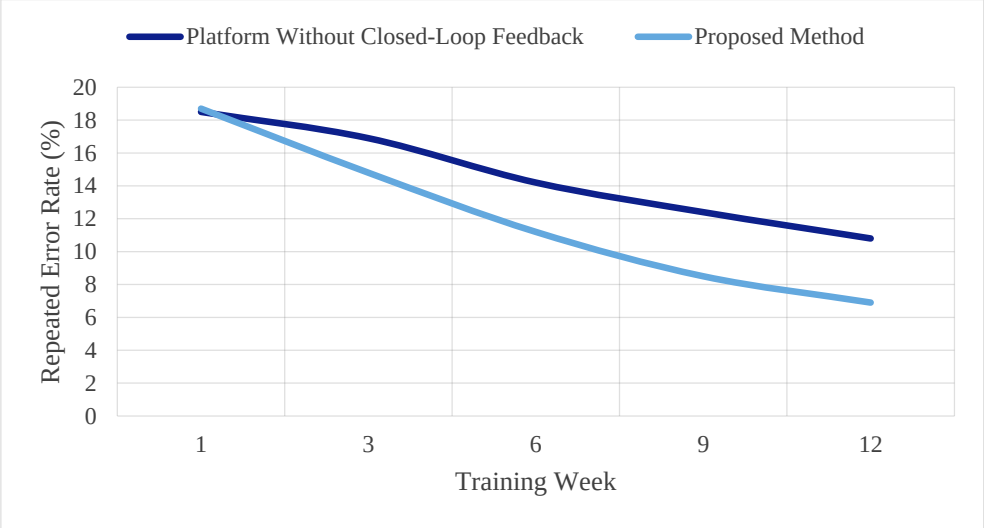

Figure 3: Curve of Error Repeatability Rate

In summary, the platform in this paper demonstrates significant advantages in four dimensions: system response, recommendation accuracy, teaching effectiveness, and feedback regulation. Among them, the recommendation accuracy has increased to 91.6%, the comprehensive performance score has increased by 12.2 points, and the error repetition rate has decreased to 6.9%. This indicates that the proposed learning model construction and intelligent regulation method can well adapt to the application requirements of multimodal, high feedback, and strong individual differences in vocal performance teaching.

6 Discussion

From the experimental results, the intelligent multimedia platform constructed in this paper does not stop at the level of digital resource presentation, but forms the whole process teaching support ability supported by multimodal resource organization, learning path generation and feedback regulation in the scene of vocal performance teaching emphasizing demonstration, imitation, deviation correction and repeated training. The results in Table 3 show that the proposed method is superior to the other comparison platforms in terms of average response time delay, recommendation accuracy and system availability, indicating that the combination of multimodal modeling and intelligent control mechanism not only improves the operating efficiency of the platform, but also enhances the stability of the platform in complex teaching tasks. At the same time, in the analysis of teaching implementation effect, this paper further takes the students corresponding to the first group of traditional multimedia teaching platform as the control group, and the students corresponding to the fifth group of proposed platform as the experimental group, and compares their training results changes under the same teaching cycle. It can be seen that the comprehensive performance of the experimental group increased from 74.6 to 86.8, with an increase of 16.4%, which was significantly higher than that of the control group of 8.4%, and the learning engagement index also increased from 0.61 to 0.83. These results indicate that the above five groups of incremental experimental designs can not only verify the technical gains at the platform level, but also reflect the promotion effect of the method in this paper on students 'singing performance and learning participation at the level of teaching effectiveness, so that the experimental design, result presentation and research conclusion form a relatively consistent logical closed loop.

Further, the introduction of personalized learning paths has directly improved the long-standing problem of "uniform tasks, uniform rhythms, and uniform feedback" in vocal teaching. Traditional platforms often place more emphasis on whether the resources are complete, but rarely deal with the issue of "what the students need to learn at the moment". After placing learning behaviors, ability deficiencies, task difficulty, and expected benefits in the same calculation link, the recommendation results are no longer simply a simple historical similarity match, but more closely resemble a dynamic arrangement oriented towards learning goals. In the experiment, the performance of this method in recommendation accuracy and convergence stability is better, indicating that the combination of behavior analysis and multi-modal features is more in line with the actual situation of obvious individual differences and inconsistent training rhythms in vocal performance teaching.

The role of the feedback loop mechanism is also worthy of attention. Many problems in vocal learning have the characteristic of continuous accumulation, such as insufficient breath support, unstable rhythm control, and loose emotional expression. If they cannot be identified and corrected in a short period of time, they will often be amplified in subsequent practice. The platform in this paper makes a joint judgment on practice results, feedback delay, and evaluation credibility, and then triggers resource rearrangement, difficulty rollback, and interaction regulation, enabling the platform to have the dynamic support capability of "while

practicing, while judging, while adjusting". Combined with Figure 3, it can be seen that the repetition error rate under this method has continuously decreased from 18.7% to 6.9%, while the platform without closed-loop regulation remained at 10.8% in the 12th week, indicating that the closed-loop mechanism does indeed improve the efficiency of problem identification and correction. Further, by combining the platform operation logs, it can be observed that the review pressure of teachers for repetitive problems has decreased after the activation of the closed-loop regulation. This indicates that the value of the feedback loop mechanism not only lies in the result level of the decrease in student error rate, but also in the fact that the platform can convert some of the process support that originally relied on teachers' repeated intervention into active regulation within the system, thereby enhancing the continuous support ability in the vocal teaching process.

Of course, this paper still has room for further improvement. Firstly, the samples mainly come from college vocal courses, and the data sources are relatively concentrated. The generalization ability in different age groups or different teaching systems still needs to be further verified. Secondly, the current model pays more attention to the joint expression of audio, video, text, and behavior logs, and the depiction of higher-level artistic features such as emotional tension and stage charisma is still relatively weak. Thirdly, although the platform has achieved dynamic recommendation and feedback regulation, its continuous adaptability in long-term teaching, cross-workpiece transfer ability, and the degree of coordination with the teacher's teaching style still need to be observed in more complex scenarios. Overall, the research in this paper is more suitable to be regarded as an intelligent implementation path for vocal performance teaching. It proves that multimodal modeling and closed-loop regulation can effectively improve the teaching support method, and also provides a basis for further promoting the digitalization and refinement of vocal teaching from a deeper level.

7 Conclusion

This paper has constructed an intelligent multimedia platform integrating multimodal resource modeling, learning behavior analysis, personalized path recommendation and feedback closed-loop regulation for vocal performance teaching. The research shows that the combined representation of audio, video, text and log can accurately depict the student's practice status, and the dynamic recommendation and rolling regulation mechanism can simultaneously improve the platform performance and teaching effect. The experimental results show that the average response delay of the platform is 0.82 seconds, the recommendation accuracy rate is 91.6%, and the system availability rate is 99.1%; the comprehensive performance of the experimental group increased by 12.2 points, the learning engagement index increased to 0.83, and the repetition error rate decreased from 18.7% to 6.9%. This indicates that the method in this paper has better adapted to the application requirements of multimodal, high feedback and strong individual differences in vocal teaching, and also provides a technical reference for the subsequent development of vocal teaching platforms towards refinement, intelligence and continuous optimization.

About the Author

ZHUO ZHANG was born in Luoyang, Henan, China, in September 1985. He is a Lecturer at Huanghe Science and Technology University, China. He received his bachelor’s degree from China Conservatory of Music, his master’s degree from Ehime University (Japan), and his Doctor’s degree from Mahasarakham University (Thailand). His research interests include

vocal performance and artistic voice science. E-mail: zhangzhuo198511@163.com

References

[1] Zhang X, Zhang J. A New Model of Vocal Music Teaching in the Context of Internet Distance Learning[J]. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 2024, 19(19): 1-12. DOI:10.4018/IJWLTT.348336.

[2] Qin H. Design of Video English Teaching System Based on Computer 5G Technology and Advanced Algorithms[J]. 2024 4th International Signal Processing, Communications and Engineering Management Conference (ISPCEM), 2024: 999-1003. DOI:10.1109/ISPCEM64498.2024.00177.

[3] Kivuti E M, Kaburu D, Ogada K O. An Interactive Multimedia Model for Developing QOE-Enhanced E-Learning Platforms[J]. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 2024, 19(19): 1-21. DOI:10.4018/IJWLTT.347662.

[4] Wu D, Shen H, Lv Z. An artificial intelligence and multimedia teaching platform based integration path of IPE and IEE in colleges and universities[J]. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 2021(2): 40.

[5] Wang M. Hybrid data clustering algorithm and interactive experience in E-learning electronic course simulation of legal education[J]. Entertainment Computing, 2025, 52(000): 100760. DOI:10.1016/j.entcom.2024.100760.

[6] Liu H, Li C. Utilizing Multimedia and Interactive Platforms in a Secondary Vocation School Politics Course: A Quasi-Experimental Study[J]. International Journal of Sociologies and Anthropologies Science Reviews, 2025, 5(1): 313-330. DOI:10.60027/ijsasr.2025.5326.

[7] Chen R, Ju H T Y, Mal N. Effect analysis of AI-assisted multimedia creation platform in college teaching[J]. Data and Metadata, 2024, 3. DOI:10.56294/dm2024615.

[8] Bajahzar A. Multimedia Educational System and its Improvement Using AI Model for a Higher Education Platform[J]. SN Computer Science, 2024(6): 5. DOI:10.1007/s42979-024-03038-2.

[9] Zhang P. Design and Implementation of English-Chinese Translation Teaching Platform Based on Deep Learning[J]. Journal of Electrical Systems, 2024, 20(3s): 1746-1755. DOI:10.52783/jes.1714.

[10] Yan S, Liu J. Design of a College English Smart Teaching Platform Based on Big Multimedia Data Technology[J]. International Journal of Web-Based Learning and Teaching Technologies, 2023, 18(2): 1-13. DOI:10.4018/IJWLTT.330676.

[11] Qian Z, Zhou T. Construction of Personalized Learning Platform Based on Intelligent Algorithm in the Context of Industry Education Integration[J]. Advances in Multimedia, 2022, 2022(Pt.7): 1.1-1.14. DOI:10.1155/2022/6042583.

[12] Liu Y. Research on Evaluation Test of Intelligent Information System in Improving the Effect of Computer Aided Education[C]. 2024 International Conference on Computers, Information Processing and Advanced Education (CIPAE), 2024: 753-757. DOI:10.1109/CIPAE64326.2024.00142.

[13] Li Z, Zhao W. The Integrated Teaching Platform of Innovation, Entrepreneurship and Moral Education in Colleges via Multimedia Network[J]. Archives des Sciences, 2024, 74(s1): 73-78. DOI:10.62227/as/74s111.

[14] Zhang T. Application of AI-based real-time gesture recognition and embedded system in the design of english major teaching[J]. Wireless Networks, 2021(6). DOI:10.1007/s11276-021-02693-0.

[15] Firmansyah F H, Sari I P, Permana F C, et al. Development of interactive learning multimedia for mathematics subjects for grade 5 elementary schools[J]. Journal of Physics: Conference Series, 2021, 1987(1): 012017 (6pp). DOI:10.1088/1742-6596/1987/1/012017.

[16] Yuanji Zhong, Jiangwei Yang, Menglong Zhang, Song Chen, Jingming Zhao. Automated Teaching Weighted Recurrent Neural Network (Atwrnn) Model: Analysis of Badminton Teaching Mode Based on Online Teaching Platform[J]. Journal of Electrical Systems, 2024, 20(3s): 1647-1658. DOI:10.52783/jes.1705.

[17] Zuhairi Z, Andayani S, Wastira J. ONLINE LEARNING MODEL TO IMPROVE THE STUDENTS' ACHIEVEMENT IN DESIGN OF INFORMATION AND COMMUNICATION TECHNOLOGY[J]. Conhecimento & Diversidade, 2024, 16(42). DOI:10.18316/rcd.v16i42.11680.

[18] Ali Q A, Sahab N M. Interactive Design of a Virtual Classroom Simulation Model Based on Multimedia Applications to Improve the Teaching and Learning Process in the Tikrit University Environment[J]. Fusion: Practice & Applications, 2023, 12(2). DOI:10.54216/FPA.120217.

[19] Qian Y .Intelligent Multimedia News Communication Platform Based on Machine Learning and Data Fusion Technology[J].Lecture Notes on Data Engineering and Communications Technologies,2023:345-354.DOI:10.1007/978-981-99-0880-6_38.

[20] Liu Q, Yang Z. The Construction of English Smart Classroom and the Innovation of Teaching Mode under the Background of Internet of Things Multimedia Communication[J]. Mob. Inf. Syst., 2021, 2021: 6398067:1-6398067:10. DOI:10.1155/2021/6398067.

[21] Huang L. An Empirical Study on the Fossilization of English Language Learning in the Context of Multimedia Network Teaching[J]. 2021. DOI:10.3233/FAIA210436.

[22] Liu L. Integration and Recommendation of Multimedia Network-Assisted English Instructional Resources Based on Association Rules Mining[J]. Mobile Information Systems, 2022, 2022(000): 10. DOI:10.1155/2022/8806525.

[23] Hu Y. Research on the Platform Construction of Multimedia Technology Education Curriculum System in Film and Television[J]. Lecture Notes of the Institute for Computer

Sciences, Social Informatics and Telecommunications Engineering, 2021: 530-536. DOI:10.1007/978-3-030-87900-6_61.