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

Personalized Feedback Mechanism for Music Teaching Integrating Artificial Intelligence Technology

Hui Wang[1,*]

[1] Zhaoqing University, School of Music, Zhaoqing City, Guangdong Province, 526000

SUMMARY: *In order to solve the problems of strong subjectivity and insufficient personalization in traditional music performance evaluation methods, this study proposes a music teaching evaluation and feedback system based on music signal processing and Convolutional Recursive Hash Network (CRNNH) model. The system first applies wavelet transform to denoise the original performance audio, and uses a pre trained FCN-5 network to extract multi-level feature maps. Then, a double-layer long short-term memory (LSTM) is further introduced to capture the temporal dependence of music signals. The test results of the MagnaTag Tune dataset show that the proposed system achieved a recognition accuracy of 91.4% in 300 recognition tasks, which is 7.2% higher than the system based on deep belief networks (DBN). When the signal-to-noise ratio (SNR) is 20 dB, the proposed system achieves an accuracy of 91.8%, and also maintains a certain advantage when the SNR is -5 dB. In addition, in terms of efficiency, the proposed system only takes 17.0 seconds to complete 300 tasks, while the DBN based system takes 22.8 seconds. In practical music teaching, the proposed system can effectively improve the accuracy of performance error recognition and provide real-time personalized feedback, which helps promote the intelligent and personalized development of music teaching.*

KEYWORDS: *Artificial intelligence; Music teaching; Personalized feedback mechanism; Intelligent evaluation; Deep Learning.*

1 Introduction

With the rapid development of Internet technology and digital audio processing technology, music signal recognition technology has been widely used in the field of education and art, providing a new and effective way to study the internal structure, acoustic characteristics and performance mode of music (Zhao, 2024; Yu and Zou, 2023). Music audio signals are not only physical representations of sound, but also information carriers that integrate art, emotion, and technology, covering multiple levels such as melody, harmony, rhythm, timbre, and dynamic changes (Rasheed et al., 2023). These signal features are not only crucial for music expression, but also help establish an automatic evaluation and feedback system for music teaching (Al Badi and Khan, 2022). With the continuous development of digital music education, there is an increasing demand for intelligent music performance evaluation tools from learners and teachers (Kakuba et al., 2022).

However, in traditional music teaching, the evaluation of student performance still largely depends on the subjective judgment and teaching experience of the teacher (Prabhavalikar et al., 2023). This subjective evaluation method often has drawbacks such as long time

*40951949@qq.com
https://doi.org/10.65102/is2026891 Published: 30 April, 2026

consumption and strong subjectivity, and is easily influenced by learners' identities, work, and teaching scenarios, lacking consistent judgment criteria. In addition, this subjective evaluation method also places high demands on teachers' professional abilities and available teaching time, making it difficult to provide one-on-one personalized guidance to all students within an effective teaching period (Weng et al., 2023). At the same time, modern learning includes both online and offline teaching methods, with an increase in learning channels, but the quality of courses varies, making it more difficult for many students to receive systematic and sustained music training (Zhang, 2024). However, receiving personalized guidance requires a significant amount of time and professional attention, and the shortage of high-quality teacher resources further exacerbates this issue (Alshamari, 2025). Therefore, in the process of music teaching, there is an urgent need for an intelligent feedback system that can expand teachers' guidance abilities. The system should be able to identify performance content, locate technical deviations, generate interpretable evaluation results, and provide useful recommendations for classroom teaching and independent practice.

Unlike regular speech signals, music signals are typically quasi periodic, containing richer timbre variations, covering a wider frequency range, and exhibiting stronger rhythm and temporal patterns (Feng, 2024). Music signals also require models and feature representations that reflect the acoustic and temporal characteristics of music performance (Shukla and Jain, 2022). In addition, audio event detection and music signal recognition typically involve several interrelated processes, including signal acquisition, preprocessing, feature extraction, representation learning, and classification modeling (Wang, 2023). Therefore, traditional speech processing methods cannot be directly transferred to music signal processing. Despite the high technical requirements for music signal processing in current related fields, the latest advances in deep learning (DL) provide a new and effective method for audio recognition and music analysis (Lu, 2023).

Currently, most DL based audio recognition models rely on neural networks and supervised learning techniques. Among them, neural networks have unique advantages in the field of audio recognition by training labeled audio samples to recognize different notes, patterns, or performance categories (Li, 2024). However, relying solely on accurate audio recognition cannot completely solve teaching problems. Because even if the system can correctly recognize the notes played by students, learners may still not understand the reasons, severity, or how to correct the errors (He and Dong, 2023). Therefore, integrating audio recognition, performance evaluation, and learning feedback is of great significance. This helps students receive more direct guidance in practice, and teachers can use the system's output to improve diagnostic efficiency and teaching organization (Ning and Jia, 2021).

In this context, this study constructed an intelligent teaching evaluation and feedback system that integrates DL and music signal processing. It can automatically recognize students' performance audio and generate personalized improvement suggestions, thereby improving the accuracy, real-time, and targeted feedback of music teaching. The core contribution of this study lies in: (1) introducing artificial intelligence into the performance evaluation process, establishing an intelligent teaching support framework that can automatically identify performance problems and provide timely guidance. (2) Developed a music signal recognition model based on deep learning, which achieves accurate recognition of performance content and quantitative evaluation of performance quality through time-frequency feature extraction and temporal modeling of performance audio. (3) Based on recognition and evaluation, targeted feedback is generated according to individual performance characteristics, which not only supports students' independent practice, but also effectively reduces the burden of repetitive evaluation on teachers, strengthening the closed-loop interaction between teaching diagnosis and learning improvement.

2 Related Work

The research on music signal recognition has gradually shifted from traditional signal processing to deep neural network driven methods, and many scholars have conducted extensive research on music signal processing and analysis techniques. For example, at the level of feature extraction and front-end processing, Tang and Zhang (2022) introduced a digital piano audio feature extraction method based on wavelet packet transform in the construction of a teaching recommendation system, enhancing the ability to capture fine audio information through spectral sensing algorithms. Zhang (2024) proposed a retrieval framework that integrates relative feature representation and deep learning for the problem of multi sentence humming retrieval. The framework combines adaptive model selection mechanism to achieve targeted classification of different beat features, and experiments have confirmed its advantages in recognition accuracy. Li and Han (2023) applied audio IoT technology to the intelligent mode of music teaching, using short-term autocorrelation methods for pitch detection and note recognition. However, due to the inherent existence of pitch estimation errors, the recognition range is mainly limited to the octave scale. To enhance the separability of beat related features, Huang (2022) designed a composite preprocessing scheme that combines short-time Fourier transform with Mel frequency cepstral analysis, effectively improving the classification performance of music signals. In addition, Pei et al. (2023) pointed out in their systematic review of instrument recognition technology in music information retrieval that matching and distance measurement algorithms based on hidden Markov models provide a reliable approach for melody similarity calculation, laying a solid foundation for the representation and modeling of music signals.

The advantages of deep neural networks in model construction and teaching applications are becoming increasingly prominent. Wang et al. (2022) used unsupervised convolutional neural networks to directly process sound signals, and the experimental results showed that its recognition performance was significantly better than the traditional Mel frequency cepstral coefficient method. He (2024) applied deep belief networks to music type recognition and automatic annotation, using raw spectra as input, combined with greedy pre training and supervised fine-tuning, achieving better classification performance than traditional methods. In the field of intelligent evaluation, He and Dong (2023) introduced fuzzy logic methods to evaluate music performances in vocal teaching, improving the objectivity of the evaluation process; Zhao (2024) explored the role of artificial intelligence in personalized music teaching quality evaluation, pointing out that intelligent evaluation can help achieve more accurate learning diagnosis. Wu et al. (2024) designed an intelligent cognitive evaluation program based on speech interaction and convolutional neural networks, which can achieve efficient recognition without additional post-processing, providing technical reference for automatic evaluation of music performances. These studies indicate that deep learning has not only made breakthroughs in recognition tasks, but is also gradually penetrating into teaching evaluation and feedback processes.

In summary, existing research has accumulated a solid technical foundation in feature extraction, deep classification modeling, and similarity matching of music signals, effectively promoting the continuous improvement of recognition performance. However, most of these works focus on improving recognition accuracy or achieving audio category recognition, with less in-depth exploration of the intrinsic relationship between recognition results and performance level evaluation, personalized teaching feedback. In real music education practice, only completing note recognition or type differentiation is far from meeting teaching needs. Therefore, there is an urgent need for an intelligent system that can automatically calibrate performance defects, provide interpretable evaluations, and generate targeted practice

recommendations based on this. Therefore, this study constructs a teaching support framework that deeply integrates music signal recognition, performance quality quantification, and personalized feedback generation, aiming to unify the above links in the same system, thereby enhancing the practical effectiveness and guiding value of intelligent technology in real-life teaching scenarios.

3 DL Based Music Teaching Evaluation and Feedback System

3.1 Application of DL

Music signals consist of fundamental and harmonic components, and for notes played by specific instruments, there is a fixed phase relationship between the fundamental and harmonic components. This periodic structure causes the entire waveform to exhibit repetitive features with the fundamental frequency as the period in the time domain, providing a theoretical basis for time-domain recognition algorithms. DL, as an important research direction in the field of machine learning (ML), is mainly based on artificial neural networks (ANN) and models complex relationships between data through multi-level nonlinear transformations. Compared to traditional ML methods, DL not only can mine the correlation between data features and tasks, but also has the ability to automatically extract higher-level abstract features from low-level features. Neural network is a highly nonlinear, ultra large scale continuous time dynamical system, whose core characteristics include continuous time nonlinear dynamic behavior, global information exchange mechanism, large-scale parallel distributed processing capability, as well as good robustness, self-learning, and associative ability.

CNN is a deep feedforward neural network specifically designed for processing grid like data, with strong applicability in image analysis, speech recognition, and natural language processing. Unlike traditional methods that rely on manually designed features, CNN can automatically learn local patterns in input data through convolutional kernels, thereby extracting different time-frequency features from audio spectrograms. This end-to-end learning method not only avoids tedious manual feature engineering, but also obtains more robust feature representations. In addition, compared to fully connected networks, CNN does not require dense connections between all neurons in adjacent layers, significantly reducing the number of model parameters, improving computational efficiency, and reducing the risk of overfitting. This structural characteristic makes CNN particularly suitable for data with spatial or time-frequency patterns, such as spectrograms generated from music signals. In the personalized feedback mechanism of music teaching, DL based models can utilize the powerful feature extraction ability of CNN to accurately capture performance related features from students' audio performances, thereby supporting more reliable recognition, evaluation, and feedback generation.

3.2 System Construction

The music teaching evaluation and feedback system proposed in this study is mainly divided into four functional layers: data acquisition, signal processing, intelligent analysis, and feedback application, which jointly realize the collection, processing, and analysis of students' performance audio, and transform it into personalized teaching feedback. At the acquisition layer, the system captures students' music performance audio through audio input devices and converts it into digital music signals for subsequent modeling. Due to the possible presence of

background noise, environmental interference, or unstable signal components in the collected raw audio, it is necessary to preprocess the audio before performing audio recognition.

At the processing level, the system processes the non-stationary characteristics of music signals through sampling, framing, and windowing operations, and improves the reliability of feature extraction by dividing continuous audio into short time periods with relatively stable acoustic characteristics. Then, the processed signal is transmitted to the analysis layer, which uses a DL based music recognition model to identify music performance content, analyze acoustic and temporal patterns, and generate quantitative evaluation results. At the application layer, these results are further visualized and transformed into personalized feedback suggestions based on the performance characteristics of each learner. The overall architecture of the proposed music teaching evaluation and feedback system is shown in Figure 1.

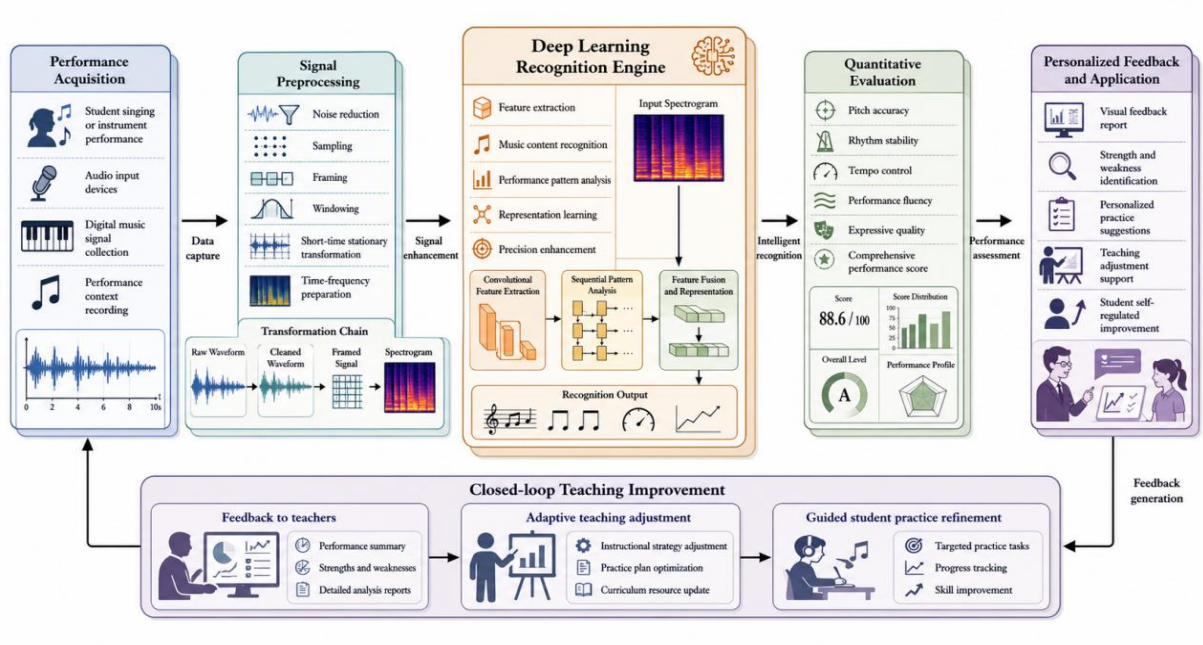

Figure 1: Overall architecture of the AI-integrated personalized feedback mechanism for music teaching.

As shown in Figure 1, the proposed system constructs a closed-loop personalized feedback mechanism for music teaching, including performance acquisition, signal preprocessing, deep learning based recognition, quantitative evaluation, and feedback application. This feedback mechanism converts raw student music performance data into interpretable evaluation results and provides personalized practical recommendations to support teaching adjustments and students' autonomous improvement.

4 DL Based Music Recognition Model

4.1 Model Building

Hash method has become one of the widely used nearest neighbor search techniques due to its fast query speed and low spatial complexity. To improve the efficiency and precision of music signal recognition, this paper combines the advantages of DL and hash learning to propose a new Convolutional Recurrent Hash Model (CRNNH), whose overall structure is shown in Figure 2. Firstly, the original music signal is preprocessed by using wavelet transform to extract

its time-frequency features, and the amplitude values of the resulting spectrogram are logarithmically transformed to obtain a more interpretable logarithmic amplitude spectrogram. Considering that music signals contain both rich spatial details and important semantic information, this paper extracts feature maps (FM) from multiple convolutional layers in a pre trained CNN network and constructs FM sequences. Among them, the first few convolutional layers preserve more spatial details, while the last layer is more inclined to capture high-level semantic features. Therefore, the FM sequence can simultaneously fuse local details and global semantic information, enhancing the representation ability of the model. Subsequently, the preprocessed logarithmic amplitude spectrum is input into a pre trained FCN-5 network, which sequentially extracts the FM outputs from each convolutional layer and integrates them through bilinear interpolation and similarity selection strategy to form a unified FM sequence.

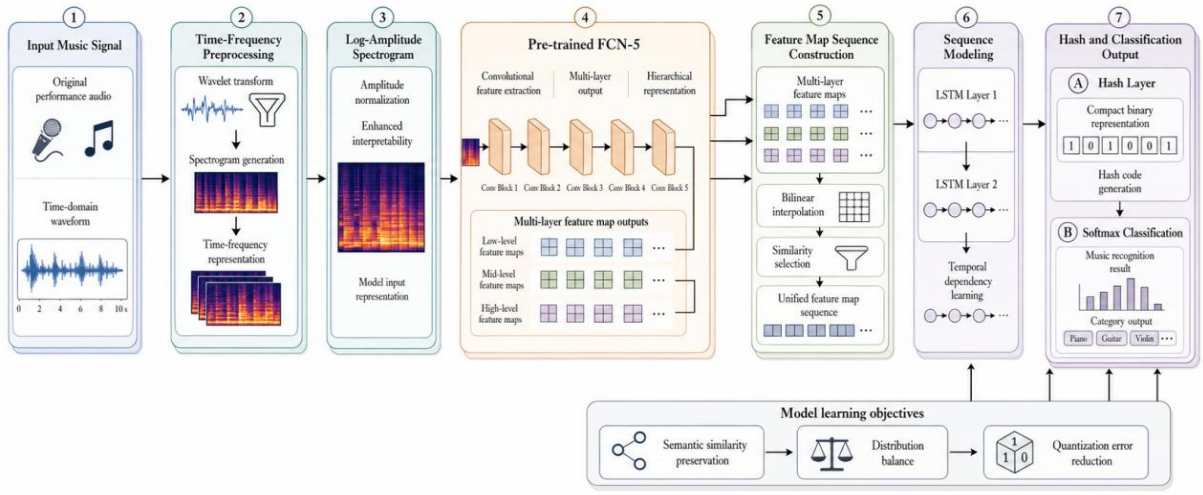

Figure 2: Overall architecture of the proposed CRNNH model for music signal recognition.

The sequence is further input into two LSTM layers and one hash layer, and finally classified and recognized through the Softmax function. The overall network structure of CRNNH includes pre trained FCN-5, two LSTM layers, a hash layer with multiple nodes, and a Softmax classification layer. Among them, the number of hidden units in the LSTM layer is set to 128. LSTM, as a special variant of RNN, is designed to solve the problems of gradient vanishing and long-term dependencies in traditional RNNs when processing long sequence data. It is widely used in tasks such as natural language processing, speech recognition, and time series prediction. It effectively solves the gradient vanishing problem of traditional RNNs in long sequence modeling by introducing cell states and three gating mechanisms (forget gate, input gate, output gate) to control the flow and retention of information. Among them, the cell state is mainly used to store historical states, the forget gate is used to control the degree of forgetting of previous state information, the input gate is used to determine whether the current input is written to the storage unit, and the output gate is used to control the amount of information output from the current state to the next layer. In addition, this article also designs a comprehensive loss function and considers the quantization error in the output binary encoding process of the hash layer to further improve the recognition performance and generalization ability of the model.

4.2 Algorithm Principle

In the process of collecting initial electronic music signals, in addition to containing the required digital sequence of music signals, interference components such as background noise are

usually mixed in. In order to achieve accurate recognition and extraction of electronic music signals, this paper uses wavelet transform algorithm to denoise the original signal. Assuming that a certain dimension of electronic music signal can be represented as a function belonging to space $L^2(R)$, its inverse Fourier transform exists in $f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)e^{j\omega t}d\omega\in L^2(R)$. Selecting an appropriate mother wavelet function $\psi(t)\in L^2(R)$, this function also satisfies the corresponding Fourier transform conditions. If the wavelet coefficients of the music signal obtained after wavelet decomposition are ω, the Fourier transform $\hat{\psi}_t$ of the mother wavelet function can be used to construct the following form of wavelet transform equation:

$$C_\psi=\int_0^\infty\frac{|\hat{\psi}(t)|^2}{|\omega|}d\omega<\infty \tag{1}$$

Among them, $\psi(\omega)$ is the frequency domain form of the wavelet mother function $\psi(t)$.

Given the input image i, it is fed into the pre trained network through forward propagation. In the forward propagation process of the network, as the network hierarchy deepens, semantic information gradually increases, while spatial details show a gradual decline trend. In addition, due to the differences in size and quantity of FM output by different convolutional layers, in order to achieve effective fusion of cross layer features, this paper adopts bilinear interpolation and similarity selection strategy to uniformly process the FM of each layer. Specifically, first adjust the FM output of each convolutional layer to the same size through bilinear interpolation; Subsequently, based on the similarity between FM, filtering is performed to retain representative FM, thereby obtaining a unified dimensional feature representation. To further clarify the internal mechanism of the proposed CRNNH model, Figure 3 illustrates how multi-level feature maps extracted from different convolutional layers are unified, filtered, and transformed into an ordered sequence representation, which is then used for temporal modeling and hash-oriented representation learning.

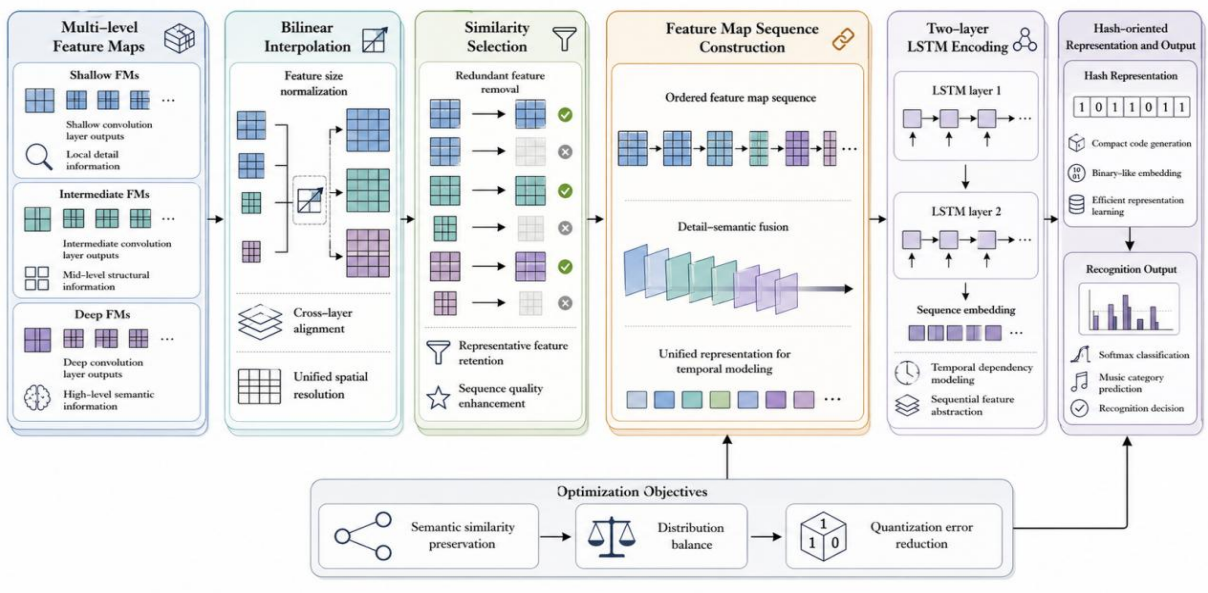

Figure 3: Mechanism of multi-level feature map sequence construction and hash-oriented representation learning in the proposed CRNNH model.

In Figure 3, the proposed method first collects feature maps of different semantic levels and spatial sizes. Subsequently, bilinear interpolation is introduced for normalization, and a similarity selection strategy is applied to remove redundant information and preserve representative feature maps. Then, the retained feature maps are sorted and input into a two-layer LSTM structure to capture temporal dependencies and sequential change patterns in the music signal. Finally, pass the optimized hash representation to the Softmax classifier and output the final recognition result.

Let x_{avg}^{s} be the average of all FMs in the s convolutional layer, and the similarity between the p FM and the average FM can be calculated using the following formula:

$$xcore_{p}^{s} = sim\left(x_{avg}^{s}, x_{p}^{s}\right) \tag{2}$$

In the formula, x_{p}^{s} is the sample image.

CNN has strong representation learning capabilities, which can efficiently and quickly extract features from audio data, accurately capturing key information in audio signals. Its core computing methods mainly include convolution and pooling operations. Among them, convolution operation is the basic step for CNN to extract local features. Through sliding filters (also known as convolution kernels), local perception is performed on the input data, and the input values of each local region are weighted and summed with the filter parameters to generate the corresponding FM. The form of convolution operation is:

$$s(t) = x(t) * w(t) = \sum_{\tau=-\infty}^{\tau=+\infty} x(\tau) w(t-\tau) \tag{3}$$

In the formula, $x(t)$ is the input feature and $w(t)$ is the feature mapping.

Input the convolutional FM sequence into the first LSTM network to generate a feature vector sequence, denoted as $H_{abstract} = \left\{\left(h_1^1, h_2^1, \cdots, h_p^1\right), \left(h_1^2, h_2^2, \cdots, h_p^2\right), \cdots \left(h_1^S, h_2^S, \cdots, h_p^S\right)\right\}$. To further explore the temporal dependencies between sequences, this paper introduces a second LSTM network, denoted as $LSTM_{encode}^{abstract}$, and takes the feature vector sequence output from the previous stage as its input. From this, the following expression can be obtained:

$$h_{end} = LSTM_{encode}^{abstract}\left(H_{abstract}, W_2, v_2\right) \tag{4}$$

Among them, h_{end} is the last hidden layer state of the second LSTM network, and W_2, v_2 are the weight matrix and bias vector of the LSTM layer, respectively.

Due to the fully connected structure between the hidden layer and hash layer of the second LSTM, the hash code can be defined as:

$$q = \phi\left(W_H^T h_{end} + \upsilon_H\right) \tag{5}$$

Among them, $W_H \in R^{H \times K}$ is the weight matrix of the hash layer, $\upsilon_H \in R^{K \times 1}$ is the corresponding bias vector, and $q = \left\{q_1, q_2, \cdots, q_K\right\}$ is the generated K bit continuous value hash code.

In the improved CNN neural network, the FM sequence is ultimately mapped to a real vector with a value range of $[-1,1]^K$, i.e. $q \in [-1,1]^K$. To convert it into a binary hash code, the following threshold function is defined in this paper:

$$b = sim(q) = sign(q_k), k = 1,2,\cdots,k \tag{6}$$

Among them, $sign(\cdot)$ is a sign function, defined as follows: when $x > 0$, $sign(X) = 1$; otherwise, -1. This function is used to map continuous values to binary outputs and is widely used in hash encoding and feature binarization tasks.

The core goal of hash methods is to generate compact and efficient binary code. However, in the optimization process, due to the constraints of discrete variables, it is usually not possible to directly use gradient descent based methods to optimize the objective function. Therefore, this study further introduced a relaxation strategy to relax the originally strict binary constraints to values within a continuous interval. After the model training is completed, these relaxed hash codes are converted into the final binary encoding through quantization operations.

Assuming that the binary code $b^{(n)}$ corresponding to image $i^{(n)}$ is used as the input of the Softmax layer, the probability of predicting the category label $y^{(n)}$ can be defined as follows:

$$p\left(y^{(n)} = m \mid b^{(n)}\right) = \frac{\exp\left(z_m\right)}{\sum_{i=1}^{M} \exp\left(z_i\right)}, m = 1,2,\cdots,M \tag{7}$$

In the formula, $z_m = w_m^T b^{(n)} + \upsilon_m$, and $b^{(n)} \in \{-1,1\}^K$, $w_m \in R^{K \times t}$ is the m weights of the softmax layer, υ_m is the m bias of softmax, and M is the number of categories in the training image.

Due to the difficulty in directly differentiating the regularization term L in its original form, a smooth surrogate function $\log(\cosh \Psi)$ is introduced for approximation (where Ψ is the difference between the predicted value and the target value, and the smaller the difference, the closer the function value is to zero and remains positive). Based on this, the regularization term can be redefined as:

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\log\left(\cosh\left(\left|q_k^{(n)}\right| - \left|b_k^{(n)}\right|\right)\right)\right) \tag{8}$$

Among them, $q_k^{(n)}$ is the element at the k position of the hash code $q^{(n)}$ corresponding to the n sample, and $b_k^{(n)}$ is the value of the binary code $b^{(n)}$ at the k position.

5 Result Analysis and Discussion

To further verify the recognition ability and robustness of the proposed system, the system based on CRNNH was compared with the system based on DBN under two verification conditions: different numbers of recognition tasks and different input signal-to-noise ratios (SNR). This experiment was conducted on the MagnaTag Tune dataset, where 27,436 samples were used for training and 9,385 samples were used for testing. The learning rate was set to 0.001, the input image size was 64×64, the momentum coefficient was 0.8, and the number of hidden units in the LSTM layer was 128.

Figure 4 shows the comparison results between our system and the DBN based system in reference [22] in terms of music signal recognition precision. As shown in Figure 4(a), when the number of recognition tasks increased from 50 to 300, the recognition accuracy of the

proposed system based on CRNNH improved from 84.1% to 91.4%, while the recognition accuracy of the system based on DBN increased from 78.2% to 84.2%. The proposed system maintained a stable advantage of approximately 5.9-7.2 percentage points, indicating that with the increase in task size, the combination of convolution feature extraction, time modeling based on LSTM, and hash representation can improve recognition stability. Figure 4(b) further evaluated the recognition robustness under different noise conditions. When the input signal-to-noise ratio increased from -5 dB to 20 dB, the accuracy of the proposed system improved from 74.8% to 91.8%, while the system based on DBN improved from 66.3% to 83.1%. Even at a low signal-to-noise ratio of -5 dB, the proposed model still achieved an 8.5 percentage point advantage, indicating that wavelet-based preprocessing and multi-level feature map construction effectively suppressed noise interference and retained useful time-frequency information. Overall, Figure 4 shows that the proposed system based on CRNNH not only achieved higher recognition accuracy in an increasing number of recognition tasks, but also demonstrated stronger robustness in noisy music signal recognition scenarios.

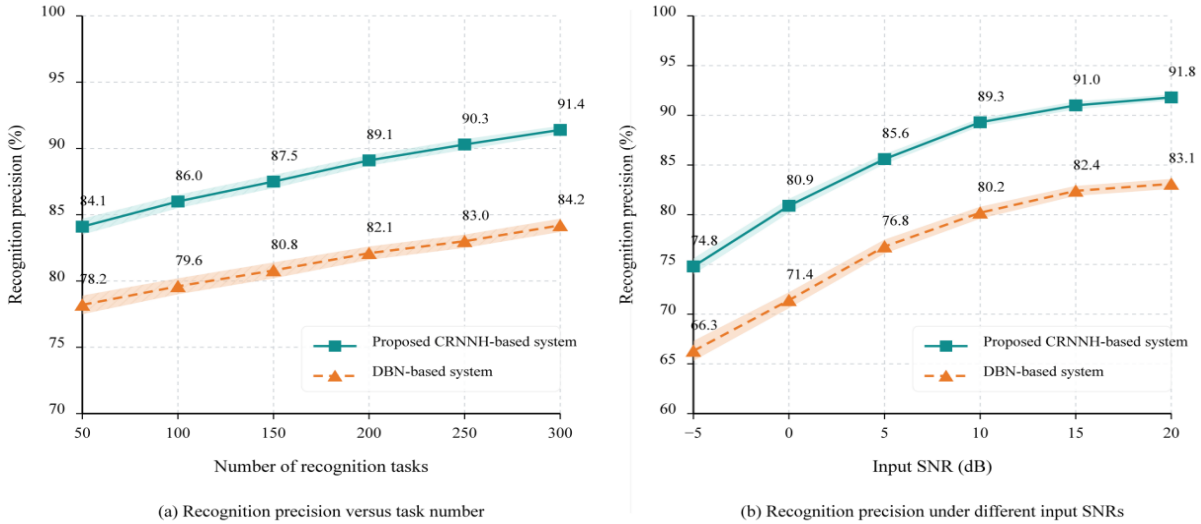

Figure 4: Recognition precision and noise-robustness validation of the proposed CRNNH-based system.

Figure 5 shows the comparison results between our system and the DBN based system in reference [22] in terms of music signal recognition response time. As shown in Figure 5(a), the total response time of both systems increased with the number of recognition tasks, which is consistent with the growing computational demand in batch recognition scenarios. However, the proposed CRNNH-based system consistently required less time than the DBN-based system. Specifically, when the number of tasks increased from 50 to 300, the total response time of the proposed system rose from 2.6 s to 11.3 s, whereas that of the DBN-based system increased from 3.5 s to 15.8 s. This means that the proposed method reduced the total response time by approximately 25.7%–28.5% across the tested task scales. Such a result indicates that the proposed model has better scalability and can maintain higher recognition efficiency when the workload increases. As shown in Figure 5(b), the average response time per task also increased with the input audio clip duration, because longer clips contain more time-frequency information and require more feature extraction and sequence modeling operations. Nevertheless, the proposed CRNNH-based system still maintained a clear efficiency advantage. When the average audio clip duration increased from 2 s to 12 s, the average response time per task of the proposed system increased from 42 ms to 118 ms, while that of the DBN-based

system increased from 61 ms to 168 ms. The proposed system therefore achieved a time reduction of about 29.8%–35.9% under different input durations. This advantage mainly comes from the efficient hierarchical feature extraction of the pre-trained FCN-5, the compact temporal modeling of the two-layer LSTM structure, and the hash-based representation mechanism, which together reduce redundant computation while preserving recognition performance. Overall, Figure 5 demonstrates that the proposed system not only improves recognition precision, but also exhibits superior real-time response capability, thereby providing stronger technical support for intelligent music teaching and personalized feedback applications.

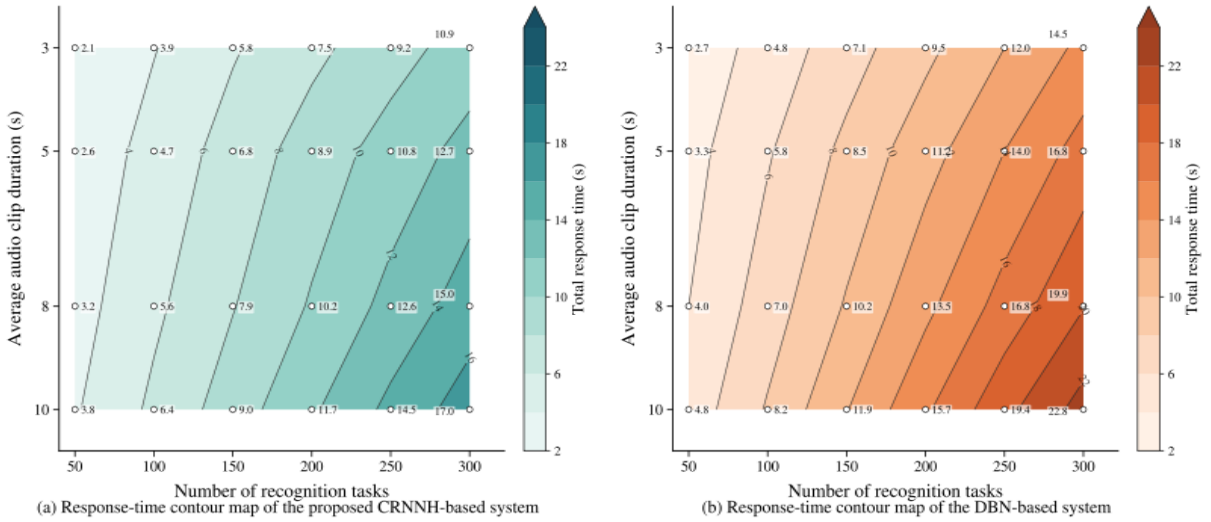

Figure 5: Response-time validation of the proposed CRNNH-based system under varying task scales and input durations.

To further evaluate the recall performance and trend stability of the proposed system, the CRNNH-based system was compared with the DBN-based system from two validation perspectives, namely different numbers of recognition tasks under different audio fragment durations and different input signal-to-noise ratios (SNR) for different tasks. The experimental setup was the same as the previous comparison experiments, including 27,436 training samples, 9,385 test samples, a learning rate of 0.001, an input image size of 64×64, a momentum coefficient of 0.8, and 128 hidden units in the LSTM layer. The corresponding recall rate trends are shown in Figure 6. As can be seen from Figure 6 (a), as the number of recognition tasks increased from 50 to 300, the recall rates of both systems gradually increased, indicating that the recognition framework remained stable under larger task scales. For the proposed CRNNH-based system, the recall rate for 3-second clips increased from 84.8% to 92.0%, and the recall rate for 8-second clips increased from 86.0% to 92.7%. In contrast, the DBN-based system increased from 79.8% to 86.6% for 3-second clips and from 80.7% to 88.0% for 8-second clips. Moreover, the performance of the 8-second clip setting was slightly better than that of the 3-second clip setting, indicating that longer audio clips provide more abundant temporal information, which is beneficial for improving recall ability. As shown in Figure 6 (b), as the input SNR increased from -5 dB to 20 dB, the recall rates of both systems increased, indicating that noise reduction and clearer signal quality contribute positively to the integrity of recognition. For the proposed CRNNH-based system, the recall rate increased from 75.2% to 91.6% under 100 tasks and from 76.8% to 92.3% under 300 tasks. In contrast, the DBN-based system increased from 68.0% to 85.4% under 100 tasks and from 69.7% to 86.1% under 300 tasks. At 20 dB, for both task scale settings, the proposed system still had a 6.2 percentage point

advantage over the DBN-based system. The results show that the proposed CRNNH-based system not only achieved higher recall rates in different task numbers and clip durations, but also maintained stronger robustness under different noise conditions, further confirming its advantages in music signal recognition and personalized feedback applications.

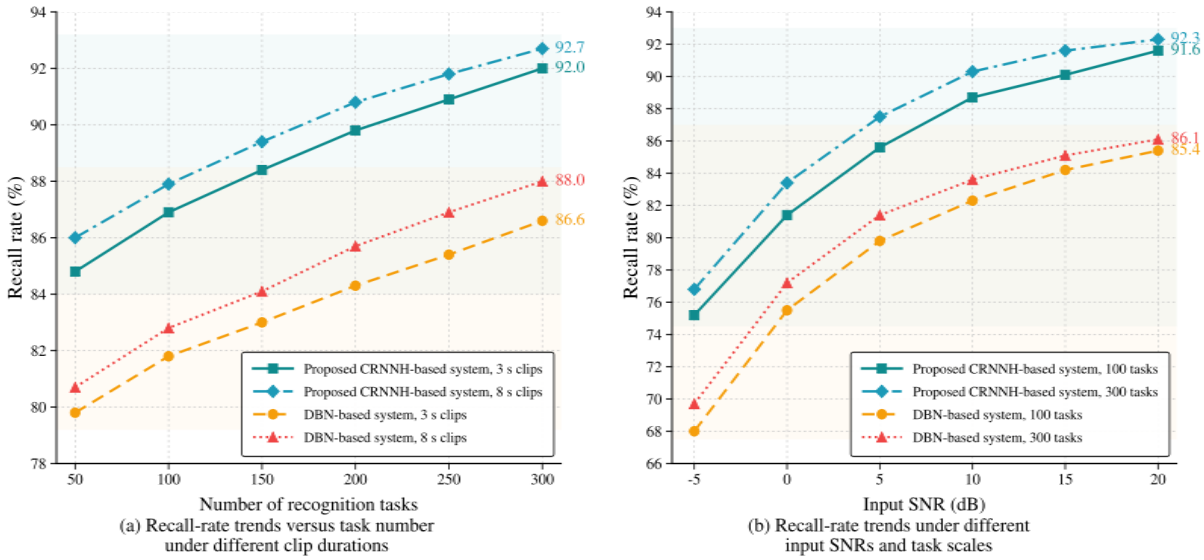

Figure 6: Recall-rate trend comparison of the proposed CRNNH-based system and the DBN-based system under varying task scales, clip durations, and input SNRs.

Figure 7 shows the comparison results of the performance level evaluation accuracy between the system proposed in this study and the DBN based system. It can be seen that as the number of tasks increases, the evaluation accuracy of both systems shows an upward trend, but this research system maintains higher accuracy under all task volumes. This indicates that in practical teaching environments, the proposed system not only can more accurately assess students' performance level, but also has stronger teaching assistance capabilities. This is mainly due to the use of CRNNH in the system, which can effectively extract relevant features from the audio played by students, such as pitch, rhythm, etc., and accurately identify problems such as note errors and unstable rhythm. These recognition results provide a reliable foundation for generating personalized feedback in the future. Compared with traditional DBN methods, the CRNNH model combines the advantages of DL and hash mechanisms, which not only improves the recognition efficiency of audio features, but also enhances the ability to distinguish different levels of performance quality, providing more solid technical support for intelligent music teaching evaluation.

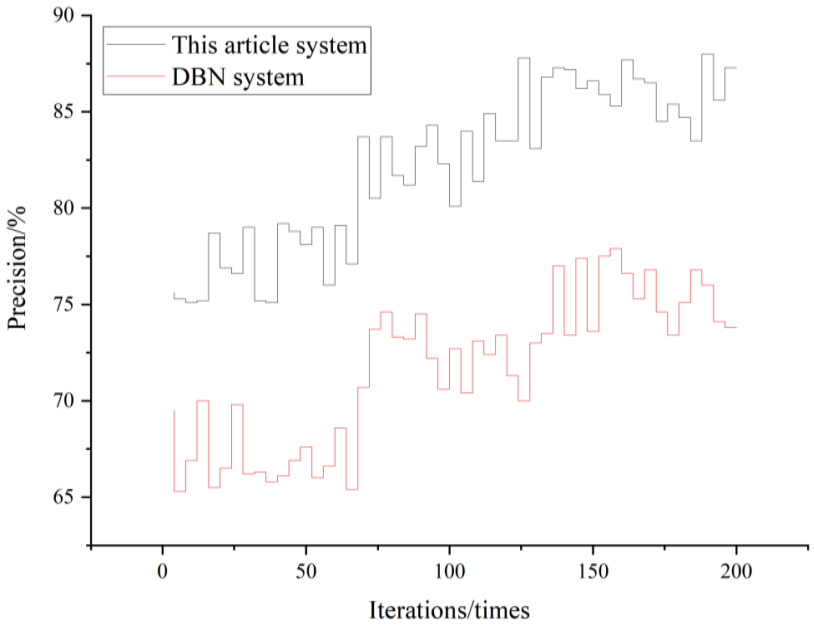

Figure 7: Comparison of evaluation precision.

To further validate the practical applicability of the proposed system, this study evaluated the role specific satisfaction trends under different task scales and the coupling relationship between performance evaluation accuracy and overall user satisfaction. Both student and teacher users participated in the satisfaction assessment after using the system for music performance recognition, evaluation, and feedback tasks. The satisfaction verification results are shown in Figure 8. As shown in Figure 8 (a), as the number of recognition tasks increases from 50 to 300, the user satisfaction of both systems gradually improves, indicating that a more stable recognition and feedback process can continuously improve user acceptance of the system. For the proposed CRNNH based system, student user satisfaction increased from 83.4% to 91.2%, while teacher user satisfaction increased from 81.8% to 90.8%. In contrast, for DBN based systems, student user satisfaction increased from 77.6% to 85.7%, and teacher user satisfaction increased from 75.9% to 84.4%. Among 300 tasks, the proposed system outperformed the DBN based system by 5.5 percentage points in terms of student users and 6.4 percentage points in terms of teacher users. As shown in Figure 8 (b), there is a clear positive coupling relationship between performance evaluation accuracy and overall user satisfaction in both systems. When the evaluation accuracy of the proposed system increased from 84.1% to 92.2%, the corresponding overall user satisfaction increased from 82.6% to 91.0%. In contrast, when the evaluation accuracy of DBN based systems increased from 77.4% to 85.3%, the overall user satisfaction increased from 76.8% to 84.2%. The results indicate that the proposed system based on CRNNH has high practical value and strong user acceptance in intelligent music teaching applications.

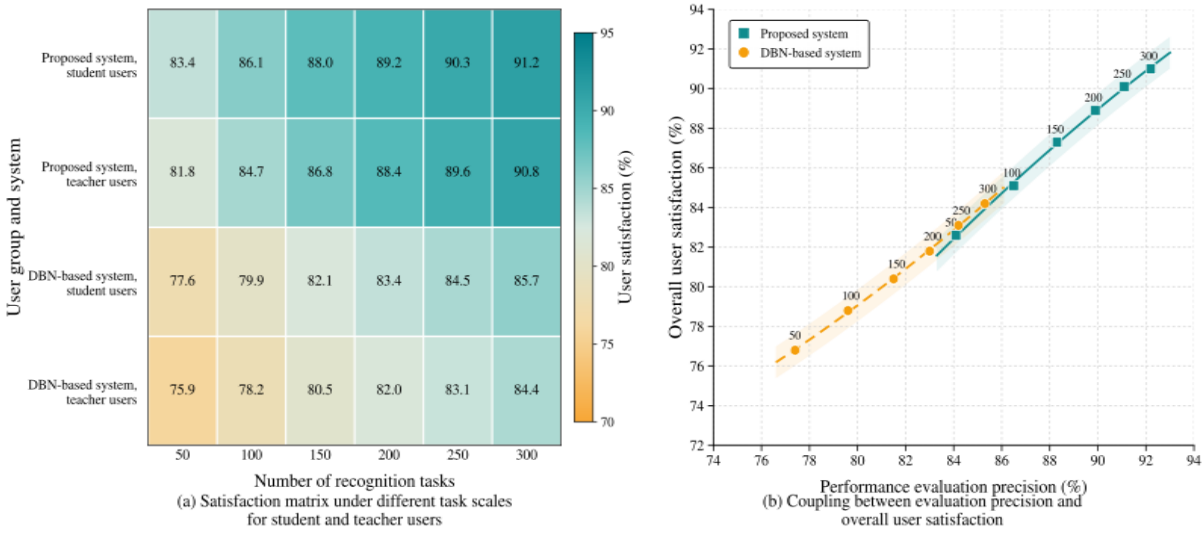

Figure 8: User-satisfaction validation of the proposed CRNNH-based system and the DBN-based system under different task scales and evaluation precisions.

6 Conclusion

In response to the problems of subjective performance evaluation, lagging feedback, and lack of personalized guidance in traditional music teaching, this study constructs an intelligent evaluation and feedback system that integrates wavelet denoising, pre trained FCN-5 multi-level feature extraction, double-layer LSTM temporal modeling, and CRNNH. On the MagnaTag Tune dataset, the proposed system achieved an accuracy of 91.4% in 300 recognition tasks, which is 7.2% higher than the DBN baseline. And when the SNR is 20 dB, the accuracy is 91.8%, and the total response time to complete tasks of the same scale is only 11.3 seconds. This is mainly due to the collaborative contribution of multiple technical mechanisms in the system. Specifically, wavelet transform suppresses noise while preserving transient details during performance, while double-layer LSTM models local rhythm changes and long-range musical sentence structures. The continuous relaxation and quantization strategy of the hash layer maintains fine-grained identification ability, thereby achieving robust recognition and real-time feedback required for computational efficiency in strong noisy backgrounds.

After embedding the system into the music teaching process, students can make autonomous adjustments based on clear error positioning. Based on this, teachers can also shift their focus from repetitive error correction to higher-level guidance such as understanding the work. However, the training data used in the experiment is mainly based on a general music annotation set, and the adaptation to different instruments, styles, and learning stages is not yet sufficient. In addition, the robustness boundary of the system in non ideal pickup environments has not been fully defined. Therefore, in future research, cross instrument and cross level domain adaptive training should be further introduced to improve the generalization ability of the model. At the same time, combining speech enhancement networks to enhance front-end noise resistance performance, and utilizing large language models to transform quantitative evaluations into clear hierarchical natural language practice prompts, in order to meet the deployment needs of daily practice scenarios.

Funding

This work was supported by the 2026 Special Project of Zhaoqing University (Party Building Research Topic): Research on High-Quality Development Paths of Integrating Lingnan Red Music into University Party Building via Digital Empowerment (No. dj202602)

About the Author

Wang Hui was born in Jiaozuo, Henan, P.R. China, in July 1982. She is of Hui ethnicity and received the master's degree from a university in P.R. China. Now, she works as an Associate Professor and the Director of the Piano Teaching and Research Office in the School of Music, Zhaoqing University, Guangdong, P.R. China. Her research interest includes the integration of Lingnan red music into university party building with digital empowerment, as well as piano teaching and research.

References

[1] Al-Badi, A., Khan, A. (2022) 'Perceptions of learners and instructors towards artificial intelligence in personalized learning', Procedia Computer Science, Vol. 201, pp. 445–451.

[2] Alshammari, A., Alzaidi, M S.. A., Alrusaini, O. (2025) 'Robust speech perception and classification-driven deep convolutional neural network with natural language processing', Alexandria Engineering Journal, Vol. 123, pp. 358–368.

[3] Feng, Y. (2024) 'Intelligent speech recognition algorithm in multimedia visual interaction via BiLSTM and attention mechanism', Neural Computing and Applications, Vol. 36, No. 5, pp. 2371–2383.

[4] He, Peidi. (2024) 'Exploration of Music Teaching Methods in Colleges and Universities in the New Media Era: Taking Sight Singing and Ear Training as an Example', News Research Guide, Vol. 15, No. 01, pp. 139–141.

[5] He, X., Dong, F. (2023) 'RETRACTED: Vocal music teaching method using fuzzy logic approach for musical performance evaluation', Journal of Intelligent & Fuzzy Systems, Vol. 54, No. 6, pp. 9289–9302.

[6] Hu, Yiling, Zhao, Zihong, Gu, Xiaoqing. (2022) 'Modeling and Evolutionary Mechanism of the Dynamic System for the Integration of AI and Education', Open Education Research, Vol. 28, No. 06, pp. 81–90.

[7] Huang, Yutong. (2022) 'The scenario application of artificial intelligence in higher education', Journal of Harbin Engineering University, Vol. 43 No,. 07, pp. 1066–1072.

[8] Kakuba, S., Poulose, A., Han, D. S. (2022) 'Deep learning-based speech emotion recognition using multi-level fusion of concurrent features', IEEE Access, Vol. 10, pp. 125538–125551.

[9] Li, L., Han, Z. (2023) 'Design and innovation of audio IoT technology using music teaching intelligent mode', Neural Computing and Applications, Vol. 35, No. 6, pp. 4383–

4396.

[10] Li, M. (2024) ‘Application of fuzzy control algorithm in music communication culture and teaching management’, Journal of Computational Methods in Science and Engineering, Vol. 24, No. 4–5, pp. 2301–2316.

[11] Lu, Y. (2023) ‘Exploring T-spherical fuzzy sets for enhanced evaluation of vocal music classroom teaching’, International Journal of Knowledge-Based and Intelligent Engineering Systems, Vol. 27, No. 3, pp. 259–271.

[12] Ning, Z., Jia, L. (2021) ‘Research on the integration and development of internet piano teaching and traditional piano teaching’, Frontiers in Art Research, Vol. 3, No. 8, pp. 77–81.

[13] Pei, Wenbin, Wang, Hailong, Liu, Lin. (2023) ‘A review of instrument recognition under music information retrieval’, Computer Engineering and Applications, Vol. 59, No. 20, pp. 34–47.

[14] Prabhavalkar, R., Hori, T., Sainath, T. N. (2023) ‘End-to-end speech recognition: A survey’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 32, pp. 325–351.

[15] Rasheed, Z., Ghwanmeh, S., Abualkishik, A. Z. (2023) ‘Harnessing artificial intelligence for personalized learning: A systematic review’, Data and Metadata, Vol. 2, pp. 146–146.

[16] Shukla, S., Jain, M. (2022) ‘Deep ganitrus algorithm for speech emotion recognition’, Journal of Intelligent & Fuzzy Systems, Vol. 43, No. 5, pp. 5353–5368.

[17] Tang, C., Zhang, J. (2022) ‘An intelligent deep learning-enabled recommendation algorithm for teaching music students’, Soft Computing, Vol. 26, No. 20, pp. 10591–10598.

[18] Wang, Xin, Lei, Jun, Li, Xiaohuan, Tang, Cheng. (2022) ‘Intelligent Interactive Learning System Based on Deep Neural Networks’, Electronic Design Engineering, Vol. 30, No. 22, pp. 73–77.

[19] Wang, Z. (2023) ‘The Effect of Intelligent Evaluation Technology on Students’ Initiative in Post-lecture Evaluation of Online Teaching’, International Journal of Emerging Technologies in Learning (iJET), Vol. 18, No. 22, pp. 88–99.

[20] Weng, Z., Qin, Z., Tao, X. (2023) ‘Deep learning enabled semantic communications with speech recognition and synthesis’, IEEE Transactions on Wireless Communications, Vol. 22, No. 9, pp. 6227–6240.

[21] Wu, Jingnan, Chen, Nan, Xia, Huanhuan. (2024) ‘Design and Application of a Voice Interaction Intelligent Cognitive Evaluation Mini Program’, China Medical Equipment, Vol. 39, No. 05, pp. 73–79+106.

[22] Yu, H., Zou, Z. (2023) ‘The music education and teaching innovation using blockchain technology supported by artificial intelligence’, International Journal of Grid and Utility Computing, Vol. 14, No. 2–3, pp. 278–296.

[23] Zhang, K. (2024) ‘Design and application of intelligent teaching system for network and new media major driven by artificial intelligence technology’, International Journal of Embedded Systems, Vol. 17, No. 1–2, pp. 150–159.

[24] Zhang, Shu. (2024) ‘The Application Status and Optimization Strategies of Artificial Intelligence in Music Education’, Journal of Wuhu Vocational and Technical College, Vol. 26, No. 02, pp. 80–82+88.

[25] Zhang, Y. (2024) ‘A Multi-sentence Music Humming Retrieval Algorithm Based on Relative Features and Deep Learning’, Scalable Computing: Practice and Experience, Vol. 25, No. 3, pp. 1799–1806.

[26] Zhao, S. (2024) ‘The role of artificial intelligence in personalized music teaching quality evaluation’, Journal of Computational Methods in Sciences and Engineering, Vol. 24, No. 6, pp. 3723–3733.