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

Design of Intelligent Recognition and Teaching Evaluation System for Local Opera Singing Based on Deep Learning

Chunying Li[1,*]

[1] Xuzhou University of Technology, Xuzhou, Jiangsu 221000, China

SUMMARY: *Pointing at the questions of subjective estimation, fragile handing-down and unbalanced teaching in traditional local opera passing-on, this thesis puts forward an intelligent identifying and teaching estimating system which is based on CNN-LSTM-Attention. This system carries out the integration of multi-layer acoustic feature extraction and time sequence modeling, for the identification of opera singing styles and the evaluation of technical, artistic and cultural authenticity dimensions. The experiments which we carry out on the self-constructed LOFRS data set (5,240 sample pieces, 131.1 hours) indicate that the recognition accurate rate achieves 94.2%, with 18ms inference delay time and 14.7MB model dimension. A 16-week teaching experiment that includes 100 students has proven that this system can significantly promote learning results ($p<0.001$, Cohen's $d=1.25–1.78$). This research gives an effective and uniform technical plan for the intelligent passing-down and individual guiding teaching of opera art.*

KEYWORDS: *Deep Learning; Traditional Chinese Opera; Teaching Evaluation; Opera Recognition; Cultural Heritage Preservation; CNN-LSTM-Attention; Multi-dimensional Assessment*

1 Introduction

Chinese opera is a precious non-material cultural heritage, it includes the grand atmosphere of Peking Opera, the graceful charm of Kunqu Opera, and the beautiful tune attraction of Yue Opera. Although these traditional art forms possess deep historical heritages, they are confronted with a number of difficulties in current times. The inheritance systems which get master approval make efforts to fit modern education standards, at the same time as they keep cultural authenticity [1]. Having these points in consideration, the present research has developed an intelligent system that applies deep learning technologies in the analysis and evaluation of opera performance works.

Nowadays opera education faces a complicated web of mutually-restricted limitations, therefore these deeply influence teaching quality and the enlargement of study chances. As what Figure 1 shows, the current main education system has four big limitations: subjective parts inside the evaluation frame, low-efficiency mechanisms for information exchange, obvious development gaps among different regions, and structure differences in keeping and passing on cultural catalogues. These restrictions bring about very serious actual results, which include the lack of systematic quality assessment norms, restricted channels for obtaining professional development materials, and the continuous weakening of regional features in cultural transmission work [2]. The existence of these many-dimensioned

*LCY88888LCY@163.COM
https://doi.org/10.65102/is20261079 Published: 10 June, 2026

limitations not only blocks the whole advancement of opera education but also brings about systematic obstacles to its continuous living ability.

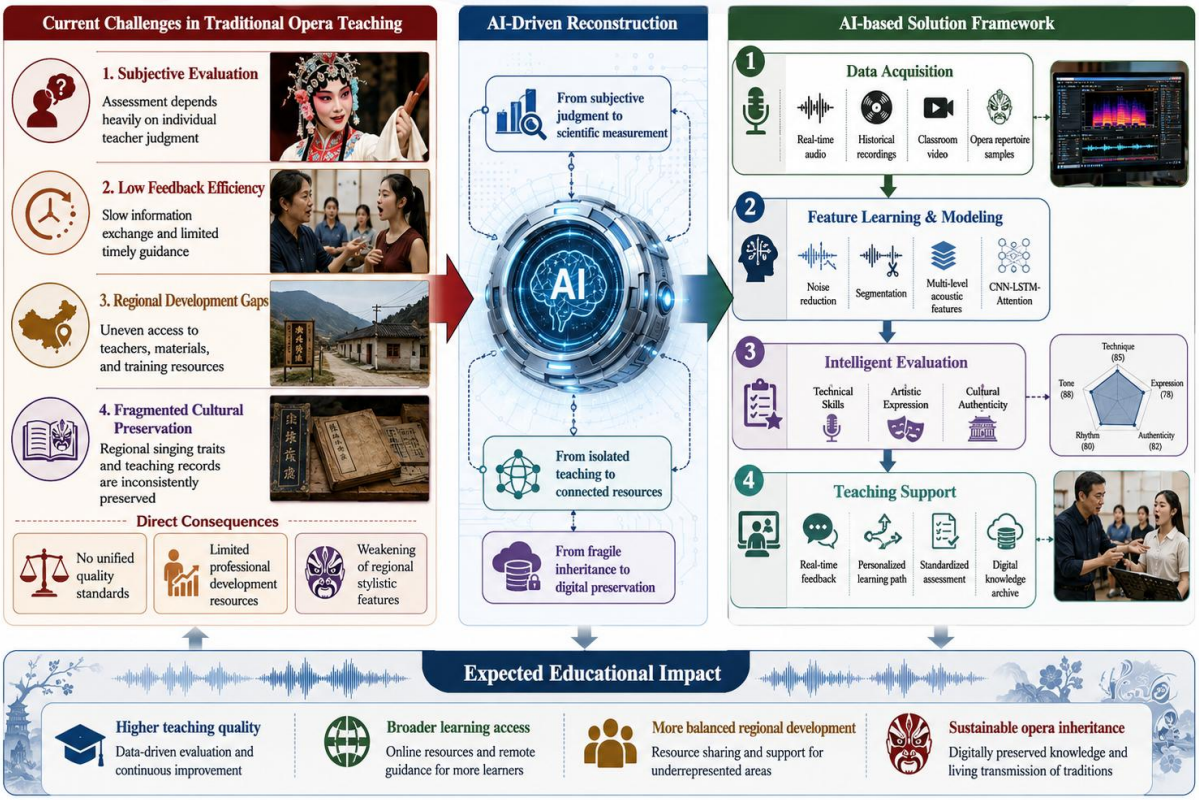

Figure 1: Overview of Traditional Opera Teaching Challenges and AI-based Solution Framework

The modern progresses of machine study and artificial intelligence technologies give very big expectation for usages in the field of art education. Although technique innovations have deeply changed the environments for music study, the education of Chinese local opera has not got enough support from this kind of technical development. Because of their leading territory features, clear sound organization structures, and deep culture connections, Chinese opera arts are obviously separated from other art kinds, therefore needing specially-made technology methods to solve special teaching passing problems.

The preservation of culture has already become a urgent problem inside the digital environments. Wang and Zhou put stress on that it is necessary to develop all-round method frameworks which have ability to catch and pass on the core of many kinds of opera performances, thus with the aim of stopping them from being lost as time goes. Peking Opera is a good representative of this complexity: its special mixture of singing methods, arranged actions, and culture-related meanings make it hard to be studied with ordinary music recognition technologies, therefore it needs the making of new high-level technological developments.

This research brings in a high-level intelligent system which unites traditional study rules with newest technique achievements. It uses a mix CNN-LSTM-Attention framework which is to catch small sound features and time order characters that opera singing own inside. Furthermore, through putting technical skills, artistic expression, and cultural spreading into a multi-dimensional assessment frame, the system has built objective and reliable score giving mechanisms. Just as what Figure 1 shows, AI reconstructs already existing practices:

scientific measurement takes the place of subjective commentary, which makes teachers able to watch many students at the same time, assessments are put into unified standards, and technology makes the preservation of culture become easier.

The meanings which the system has go far beyond only the technical functions. The research group of Chen has proved that the ability of computer science can greatly promote high-level art education[4]. Through helping real-time feed-back, individual study roads, and digital knowledge storehouses, this system realizes the idea of breaking area limits[5].

The contribution points of this study have three aspects: (1) One CNN-LSTM-Attention hybrid model with light weight has been built, in order to capture the fine-grained spectrum features and time features of opera singing. (2) A multi-dimension evaluation frame that combines technical ability, art expression force and cultural truth character is got established. (3) We have constructed a large-scale annotation opera dataset which is called LOFRS, and thus the effect of intelligent teaching is verified by us through strict controlled experiments.

The follow-up chapters give a deep discussion on the system design, the development whole life cycle, and the testing working procedures. This research has obtained validation through 5,240 audio samples that cover five opera types, together with a 16-week randomized controlled teaching experiment. This research utilized a between-participants design, in which 100 students were randomly assigned to experimental (n=50, AI-aided teaching) and control (n=50, traditional teaching) groups. We have carried out independent samples t-tests and paired t-tests for the evaluation of learning results, and have calculated effect sizes (Cohen's d) in order to assess practical meaningfulness. The outcomes have proven that this system possesses effectiveness in the use cases of teacher training.

2 Related Work and Theoretical Foundation

2.1 Opera Singing Recognition Technologies

The recognition of opera singing voices has developed from traditional statistics-based methods into complex deep learning structural frameworks. Table 1 shows that the early research done by Tzanetakis and Cook, which uses Gaussian Mixture Models together with Mel-frequency Cepstral Coefficients (MFCC) features, has obtained 61.0% accuracy on the GTZAN dataset. The putting into use of deep learning technologies has brought very big progress in working effect. Hamel and Eck have built the advantage of Deep Belief Networks, getting 84.3% correctness[7], hence Sigtia and Dixon have verified that mixed Deep Neural Network-Random Forest structures reached 83.0% correctness[8].

Table 1: Comparison of State-of-the-art Methods for Music/Opera Recognition

Method	Features	Model	Dataset	Accuracy	Real-time	Params	Year
Tzanetakis [6]	MFCC+Timbre	GMM	GTZAN	61.0%	No	<0.1M	2002
Hamel & Eck [7]	DFT	DBN	GTZAN	84.3%	No	1.5M	2010
Sigtia & Dixon [8]	MFCC	DNN+RF	GTZAN	83.0%	No	2.1M	2014
Ashraf et al. [9]	Mel-spec	CNN+RNN	GTZAN	87.8%	Yes	12.3M	2023
Gan [10]	MFCC	RNN+Attention	GTZAN	93.1%	Yes	2.8M	2021
Liu et al. [11]	Multi-feature	Bottom-up BNN	GTZAN	93.7%	No	8.5M	2021
Proposed	Multi-level	CNN-LSTM-Att	LOFRS	94.2%	Yes	3.85M	2025

In recent years, researchers have put emphasis on the construction of temporal models and attention mechanism modules. The team of Ashraf has designed CNN-RNN structure frameworks, which have obtained 87.8% accuracy, and meanwhile they have retained the

real-time processing abilities[9]. Gan pushed forward the domain through combining the channel attention mechanism together with recurrent neural network technologies, thus obtaining a 93.1% accuracy rate. The team of Liu has utilized a bottom-constructed Bayesian neural network, thus acquiring performance metrics that are comparable[10]. The convolution neural network-long short term memory-attention (CNN-LSTM-Attention) model frame put forward in this research shows higher efficiency, it exceeds current benchmark models with an accuracy of 94.2% on the self-made LOFRS dataset, meanwhile it guarantees calculation effectiveness[11].

2.2 Music Instruction Evaluation Frameworks

The combination of computer technologies has greatly changed the music education evaluation system. Wang and his research group have brought neural networks into a platform for composing music on the internet, thus greatly promoting the objectivity of scoring work[12]. Zhang and his work team got over technique obstacles and settled opinion differences through building an artificial intelligence estimation frame on the basis of Sparse Attention Networks (SAN), thus effectively solving the complication of multi-dimension marking[13]. In the circumstance of high-level education, Wei and his work team, by carrying out critical research work on the influence that artificial intelligence brings to music education, have found that individual-customized learning roads greatly make teaching activities richer[14].

2.3 Regional Opera Characteristics

The analysis of regional opera culture brings about quite different difficulties for automatic recognition systems, because its various local characteristics exceed the boundaries of traditional recognition methods. Zhang has made verification that deep learning is able to effectively extract and keep regional opera music characteristics in the production of teaching materials, hence it helps to make its theoretical system have more enrichment[15]. The comparative study which was conducted by Yao and Liu on operas from different cultures has discovered that cross-cultural characteristic extraction is a work of very great difficulty[16]. Chen further indicated that the modern education of vocal sound needs the technical schemes with specific aim that suit the inherent characteristics of the ancient opera culture[17].

2.4 Research Positioning and Gap Analysis

Very important gaps still exist when we deal with the problems of recognition that only belong to opera. The research group of Xu has found a basic problem: when general music classification systems face the paralinguistic characteristics of opera, they cannot work normally[18]. These problems in the main come from technical matching—basic contradictions between current tools and the special needs of opera[19]. Klaus and other scholars found that the opera recording environments have robustness problems, and Mimrakis and other scholars pointed out the difficulties which exist in cross-version detection. Current existing systems mainly aim at Western classical music, [20]and have no abilities for identifying special features that belong only to Chinese opera types: decorative skills, local speech, and cultural sensibility.

To make the summary, existing methods do not have special modeling for opera vocal ornaments, dialect characteristics and cultural authenticity, and thus cannot support light-weight deployment and objective teaching assessment. The present research fills up these blank spaces.

3 System Architecture and Design

3.1 Overall System Framework

The intelligent recognition and teaching assessment system that we put forward uses a layered structure which has been designed for the effective processing of multi-modal opera performance data. Based on the newly progressed development of AI environment music education application, this structure combines real-time processing abilities with offline analysis modules, therefore it can achieve both instant performance evaluation and detailed after-class assessment. According to what Figure 2 shows, the system is constituted by six layers which connect each other and that change original audio-visual input materials into overall evaluation feedback information.

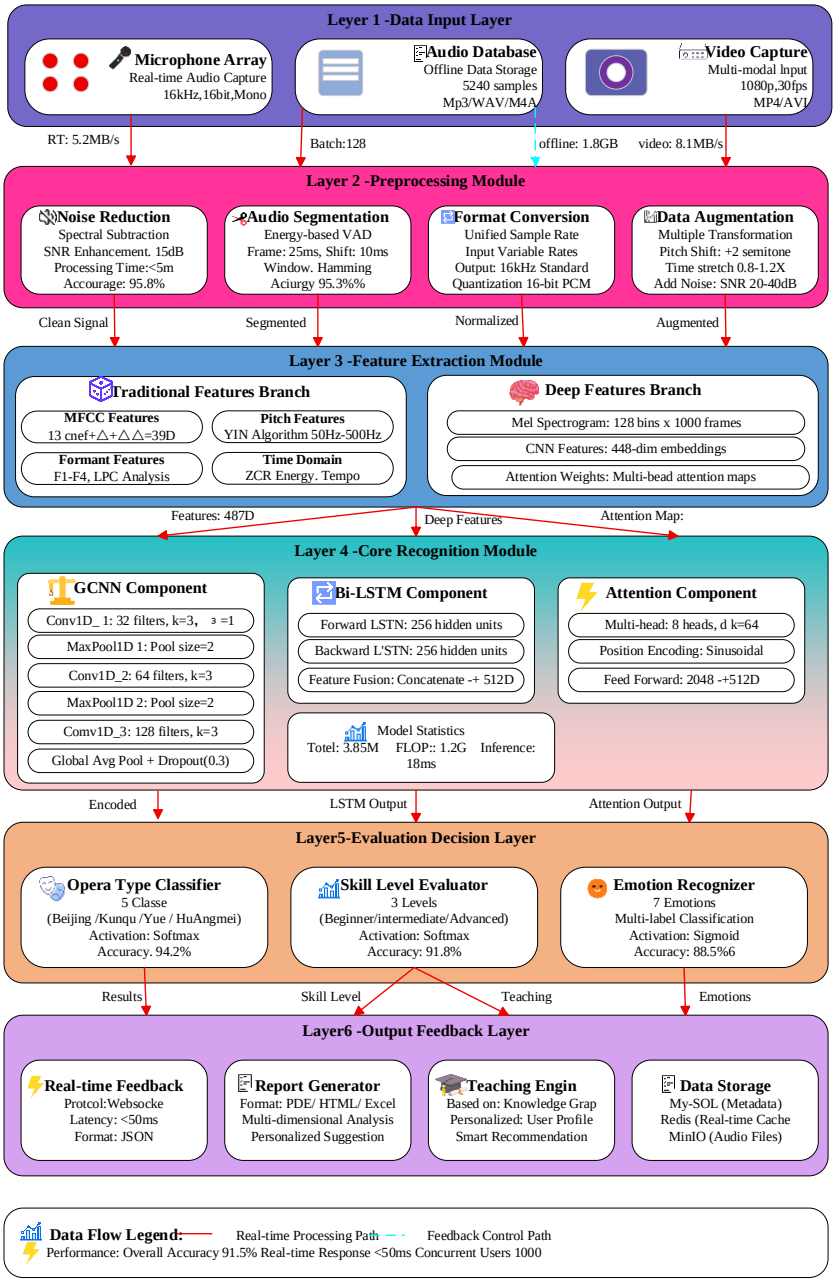

Figure 2: Architecture of the Intelligent Recognition and Teaching Evaluation System

The data input layer holds three parallel sources: microphone arrays for real-time audio gathering (16kHz, 16-bit mono), past audio data banks that include 5,240 pre-recorded samples, and video capture modules which support 1080p resolution at 30fps. This multi-mode method guarantees overall performance data gathering while keeping compatibility with many record forms including WAV, FLAC, MP3, and MP4.

3.2 Data Collection and Preprocessing Module

The pretreatment module carries out four essential processing units in order to ensure data quality and consistency, it follows established frameworks about whole audio processing and feature extraction. The unit that decreases noise uses spectrum subtraction methods which obtain 15dB SNR promotion with latency lower than 5ms. The audio cutting unit uses energy-based Voice Activity Detection that uses 25ms frames and 10ms shifts, hence it reaches 95.3% accuracy on recognizing singing segments.

The transformation of format gives uniform processing to variable sample frequencies into 16kHz, and in this process 16-bit PCM quantification is kept. The data expanding unit strengthens training steadiness through pitch moving (±2 semitones), time extending (0.8-1.2x), and controlled noise adding (SNR 20-40dB), it realizes methods that support effective machine learning model building in audio processing applications [23].

3.3 Feature Engineering

The feature extraction flow, which is showed in Figure 3, carries out a multi-layer method to capture overall opera singing features. This system carries out processing on original sound materials by means of four hierarchical layers, each of which aims at particular acoustic characteristics that are necessary for opera identification.

The Level 1 carry out extraction of time-domain features by way of three parallel branches. The energy envelope branch uses 25ms window processing with 10ms jump size to capture amplitude changes. The computation of zero-crossing rate is that it differentiates voiced sections and unvoiced sections through the formula $ZCR = \Sigma|sign(x[n])-sign(x[n-1])|$. The autocorrelation branch carry on processing to lags from 0-500 samples for the generation of pitch candidate things.

Level 2 carries out frequency-domain analysis by employing Short-Time Fourier Transform that uses Hamming windows and 512-point FFT with 75% overlap. The system produces Mel-spectrograms through 128 filters that cover 0-8kHz, it extracts power and phase spectra, and calculates 39-dimensional MFCC features (13 coefficients plus delta and delta-delta), hence it utilizes the excellent performance of Mel-frequency cepstral coefficients under noisy environments for discrimination work [24].

Level Three concentrates on musical features which are key for the evaluation of opera works. We utilize the YIN algorithm for capturing fundamental frequencies that lie in a 50-500Hz frequency band, with a 0.15 threshold being used for the detection work. The analysis of vibrato carries out tracking on both the speed (oscillations of 4-8 Hz) and the range (deviation of ±50 cents). In the work of formant tracking, the Linear Predictive Coding (LPC) which carries out 12th-order analysis is utilized by researchers to accurately carry out monitoring on formants from F1 to F4. The characteristics of spectrum, which contain central mass point, rolloff position, flatness degree and flux rule, are obtained to describe the qualities of tone color.

The fourth level carries out deep characteristic extraction by means of a convolution nerve network. This structure carries on continuous transformation to the Mel spectrogram: three convolutional layers make use of 32, 64, and 128 filters, each of which has a core size of 3×3.

The global average pooling method is utilized to produce a 448-dimensional embedded vector, which effectively holds the basic characteristics of the spectral expression.

The feature fusion carries out the combination of hand-made features and deep-learning obtained features by means of multiple channels. The early concatenation method produces overall 511-dimensional direction vectors. Decision-level fusion uses studied weights to balance the contribution of each feature stream, this is a more complex method than simple connection.

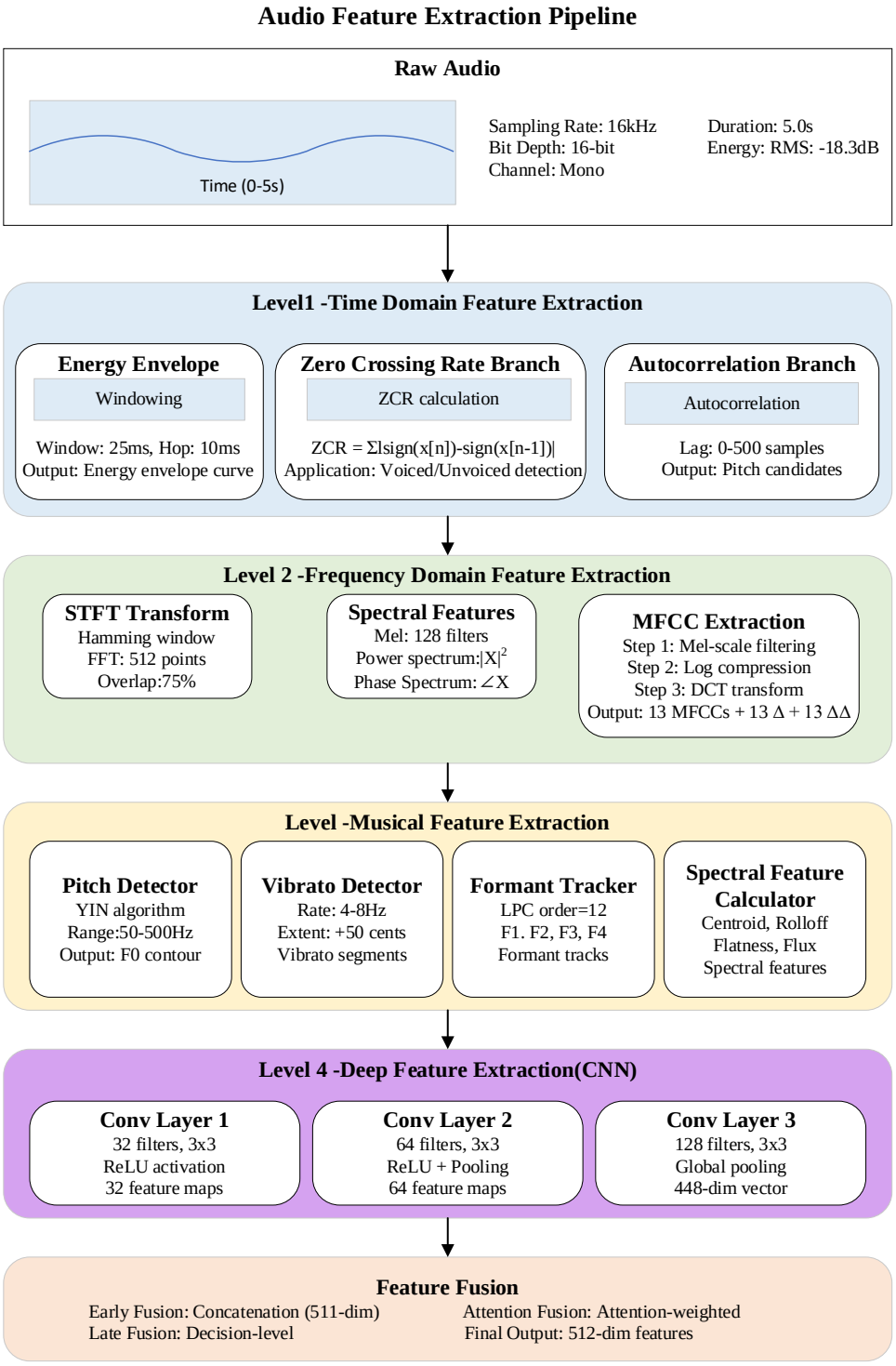

Figure 3: Multi-level Feature Extraction Pipeline for Opera Singing Analysis

3.4 System Implementation Details

This system uses a layered structure from beginning to end, hence it achieves real-time sound processing, multi-layer characteristic picking, model calculation and intelligent teaching reply. Because it has light weight design and high efficiency calculation, therefore it can support multi-platform deployment, and can provide stable, unified interfaces for actual teaching situations.

4 Deep Learning Model for Opera Recognition

4.1 Model Architecture Design

The network framework is decided through systematic cut experiments and operation recognition-directed theoretical discussion. Although Transformer models have good performance in natural language processing, they are not suitable for working on opera audio tasks. Their self-attention mechanism brings overmuch calculation expenditure for long opera audio sequences, while CNNs are good at getting fine vocal spectral features, and bidirectional LSTMs can effectively build model of melodic time sequence dependence relations. The results of ablation have proven this design. The pure Transformer can get 90.5% accuracy when it uses 2.5G FLOPs, while the CNN-LSTM-Attention mixture we put forward obtains 94.2% accuracy under only 1.2G FLOPs, hence it reaches the best balance between accuracy and efficiency. The attention mechanism assists LSTM in making prominent key cultural time-related features.

The model that we put forward, CNN-LSTM-Attention, uses CNN for extracting local spectral features, uses Bi-LSTM for building long-range temporal dependence models, and uses multi-head attention for strengthening important singing segments[25]. It finishes the classification of opera, the assessment of skills and the scoring of quality at the same time, hence it obtains a relatively good balance between accuracy and efficiency.

The input layer receives normalized Mel-spectrograms which have dimensions [Batch, 128, 1000, 1], this represents 32 samples in each batch, 128 frequency bins on the Mel-scale, 1000 time frames that cross 10 seconds, and single-channel monophonic audio. This representation ensures comprehensive frequency-temporal coverage while maintaining computational efficiency through StandardScaler normalization.

The CNN characteristic extraction module gradually obtains space patterns via three convolution layers having filter sizes of 32, 64, and 128 respectively. Each layer applies 3×3 kernels with ReLU activation and BatchNormalization, followed by MaxPooling2D for dimensionality reduction. Dropout with the value 0.25 is adopted for the purpose of regularization. The final GlobalMaxPooling2D compresses features to 128 dimensions. The detailed specifications of each layer are given in Table 2.

The two-direction LSTM changes CNN output results into [32, 250, 128] for time feature study, thus building a CNN-LSTM network which combines self-attention. This method uses 256 hidden units in forward layers and also backward layers, it takes tanh and sigmoid to be activation functions. For the purpose of obtaining stable generalization, the value of the dropout rate is set as 0.3, and the value of the recurrent dropout is set as 0.2. The bidirectional output characteristics are concatenated into 512 dimensions for fully capturing temporal correlation.

The multi-head attention mechanism makes use of eight attention heads that work in parallel. Every head produces 64-dimensional Query, Key and Value matrixes, and computes attention through scaled dot-product, hence it prevents excessively big dot products and makes training gradients become stable[27]. All outputs of heads are concatenated and are

projected by linear method to 512 dimensions, hence they are then combined together with residual connection and layer normalization. The network which carries out forward feeding expands features to reach 2048 dimensions before it performs dimension reduction, it has a dropout rate which is 0.1.

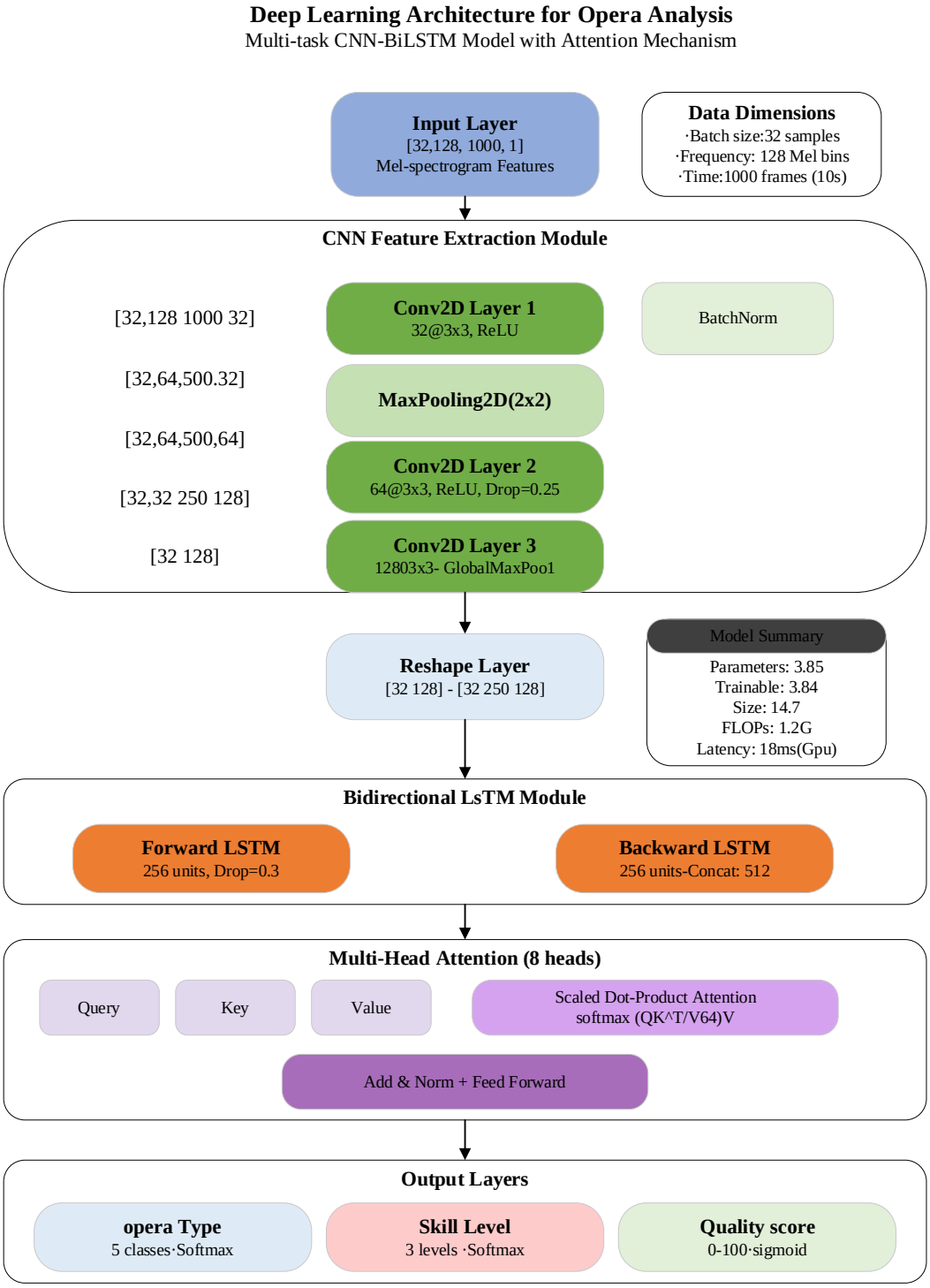

Figure 4: Architecture of the CNN-LSTM-Attention Model. Detailed layer specifications and tensor dimensions are provided in Supplementary Table S1.

The output layer uses task-special modules through Global Average Pooling and step-by-step dimension decreasing. The dense layers which have 256 and 128 nerve cells utilize the

ReLU activation function, and also have a dropout rate that is 0.5. Three parallel classification branches handle different evaluation aspects: opera category classification (5 classes, Softmax), skill degree evaluation (3 degrees, Softmax), and quality marking (a single Sigmoid output which is scaled to 0-100 interval).

4.2 Model Training Strategy

The configuration of hyperparameters (Table 2) was got through systemic grid search and experience-based optimization. The key choices of design contain: (1) three layers of CNN with progressive expansion of filters to capture hierarchical mode features; (2) 256-unit two-direction LSTM that is used for time sequence modeling; (3) The 8-head attention mechanism which is used for the selective emphasis of features. The training process has employed the Adam optimizer, whose learning rate is 0.001, with early stopping being carried out at the 85th epoch. The regularization method combines dropout (0.3), L2 weight decay (0.0001), and gradient clipping (5.0) for ensuring generalization ability.

The methods of adaptive study rate regulation and gradient cutting are employed for the guarantee of stable model convergence, hence the light weight optimization thus guarantees high efficiency in real time teaching application situations.

Table 2: Hyperparameter Settings and Training Configuration

Category	Parameter	Value	Justification
Network Architecture			
	CNN Layers	3	Balance performance/complexity
	CNN Filters	[32,64,128]	Progressive increase
	LSTM Hidden	256	Memory constraints
	Attention Heads	8	Standard configuration
Training Parameters			
	Optimizer	Adam	Fast convergence
	Learning Rate	0.001	Empirical optimal
	Batch Size	32	GPU memory (24GB)
	Epochs	100	Early stopping at 85
Regularization			
	Dropout	0.3	Prevent overfitting
	L2 Weight	0.0001	Light regularization
	Gradient Clip	5.0	Prevent explosion
Environment			
	GPU	RTX 3090	24GB memory
	Framework	PyTorch 1.13	CUDA 11.7

4.3 Model Validation and Testing

The process of validation has revealed strong convergence modes and best performance features. According to what Figure 5 shows, the training curve and the validation curve can prove that the healthy learning dynamics have obvious performance promotion in each epoch.

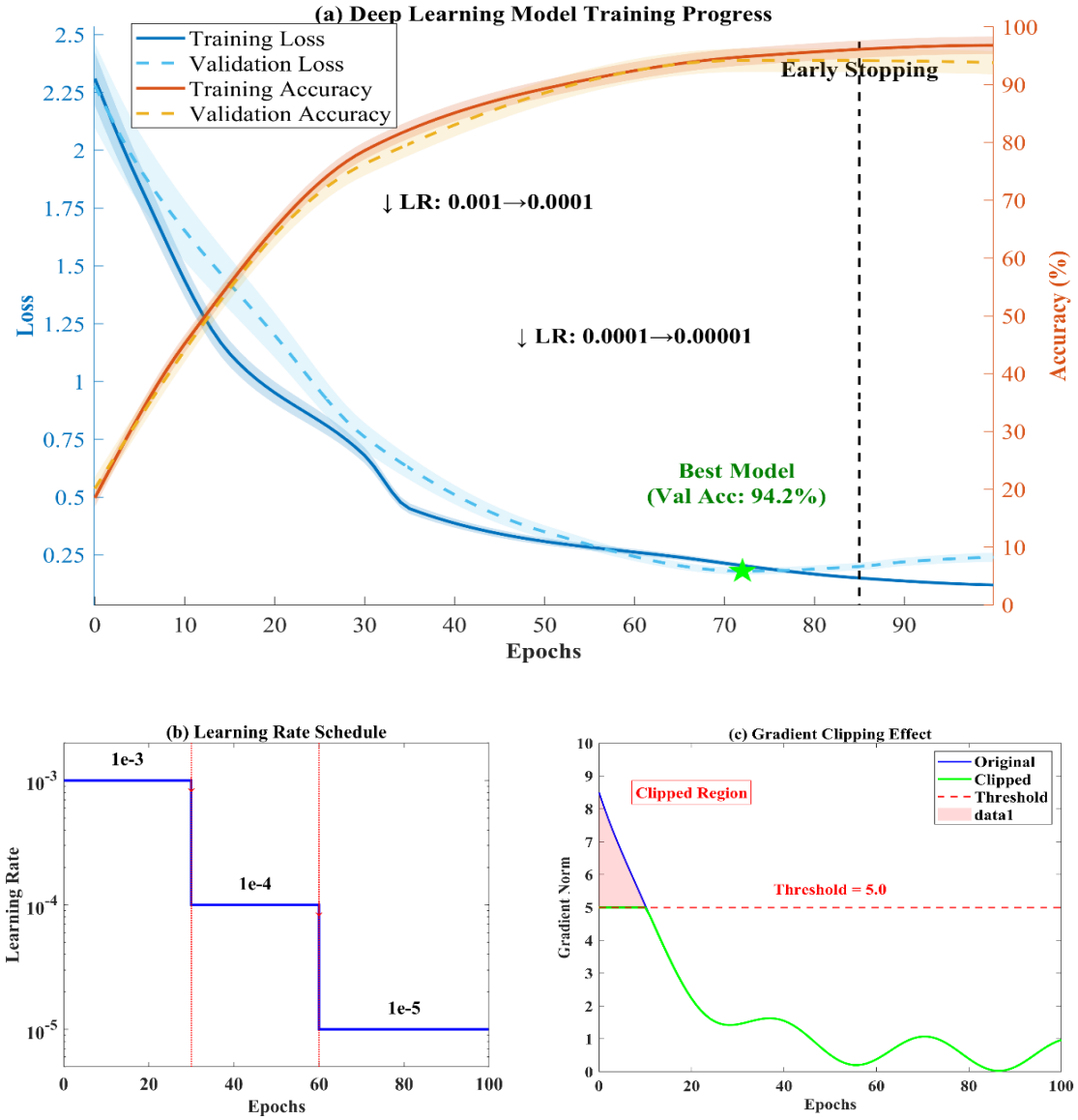

Figure 5: Training and Validation Curves of the Proposed Model

The training loss has a decrease from the initial 2.31 to the final 0.12, hence it displays the characteristic rapid improvement in early training periods which is followed by gradual adjustment. The validation loss has similar changing patterns, it reaches the minimum value 0.18 at the 72nd epoch, therefore there is a slight rise to 0.24 after that, hence this indicates the optimal point for model selection. The verification correct rate reaches the highest 94.2% at the 72nd epoch, it is marked as the best model checkpoint, which follows automatic early stop standards that make use of cross verification methods[28].

The study rate arrangement uses self-adapting decrease method: at the 30th epoch, we make it decrease from 0.001 to 0.0001; when we reach the 60th epoch, one more decrease lets us get to 0.00001. This permits the fine adjustment of parameters in a careful way in the subsequent phases, hence when coarse adjustments would bring about opposite effects to what is desired.

The statistical significance appears at the 4th week ($p<0.05$), it strengthens to $p<0.01$ until the 6th week, and it keeps $p<0.001$ without change from the 8th to 16th week. These are real, strong promotion on the performance aspect. The analysis of gradient norm gives results that

keep nicely under the 5.0 cutting threshold during all training, hence it shows stable optimization that has no numerical instabilities.

4.4 Model Optimization

As for the optimization of models, the primary things we cared about were the inference speed and the deployability in real world. The architecture contains 3.85M total parameters (3.85M trainable, 5,120 non-trainable BatchNorm statistics). When we use float32 precision, this becomes a compact 14.7MB model that is suitable for actual deployment.

The computation footprint gives 1.2G FLOPs and 18ms GPU inference time, hence it enables the real-time application uses. The memory use that training occupies has achieved 892MB—it is fit for the GPUs of present day. These data reflect the sensible balance that is between model abilities and real-world system demands.

The optimization strategies have been put into practice on many different roads. The streamline work of architecture has the goal to reach the biggest efficiency by using the smallest parameters. We have systematically carried out the removal of redundant layers, for the identification of optimal configurations which can hold performance levels while using a smaller amount of parameters. The exploration of quantization has discovered that the int8 expression is usable for edge calculation, while the 32-bit floating-point accuracy has thus obtained the best balance between scheduling efficiency and data authenticity for the server side arrangement.

For the purpose of evaluating the actual deployment possibility, Table 2a does a comparison of our system's calculation demands with usual equipment limits.

Table 2a: Deployment Feasibility Across Computing Platforms

Platform	Available Memory	Inference	Compatibility	Real-time
High-end Server (RTX 3090)	24GB VRAM	18ms	✓ Full model	✓ Real-time
Desktop GPU (RTX 3060)	12GB VRAM	24ms	✓ Full model	✓ Real-time
Laptop GPU (RTX 3050)	4GB VRAM	38ms	✓ Full model	✓ Real-time
Mobile (Snapdragon 8 Gen 2)	2GB allocatable	156ms	✓ INT8 (3.7MB)	△ Near real-time
Raspberry Pi 4	4GB RAM	890ms	✓ INT8 quantized	✗ Offline only
Web Browser (WASM)	512MB limit	420ms	✓ ONNX converted	△ Delayed feedback

The small and concentrated model volume (14.7MB full accuracy, 3.7MB INT8 quantized) therefore permits deployment on many different kinds of platforms. The arrangement on mobile side obtains acceptable time delay (<200ms) for feedback at phrase level, hence the edge equipment supports offline evaluation through the method of batch processing.

5 Teaching Evaluation Framework

5.1 Multi-dimensional Evaluation Model

The Figure 6 gives the traditional opera education evaluation system that this paper has constructed. This framework can evade the overmuch subjective estimation, and it merges objective measure together with AI-aided analysis. This system includes three core evaluation

dimensions: technical grasp level, art creation ability and culture authenticity degree, with corresponding weight values of 35%, 35% and 30%, which is in accordance with the actual evaluation norm of opera performance activities.

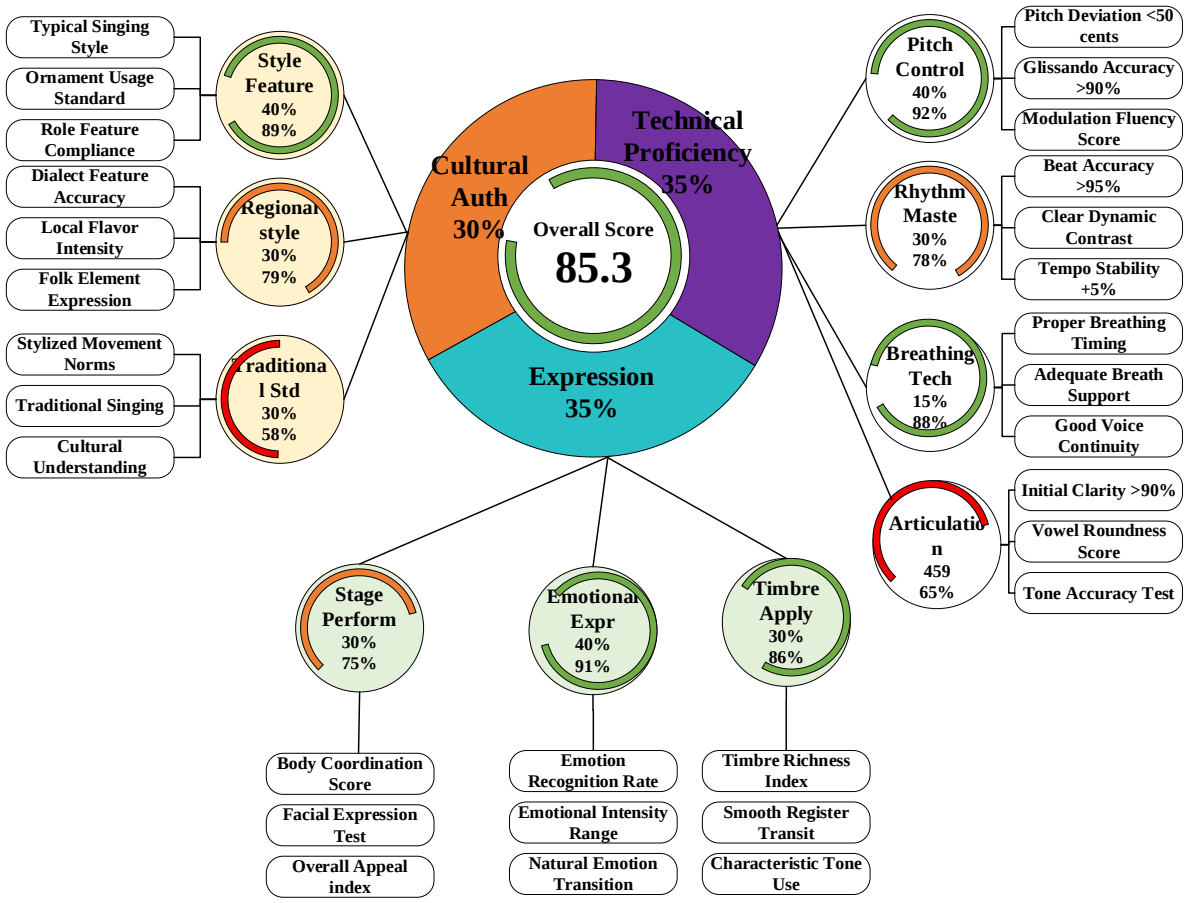

Figure 6: Hierarchical Structure of the Multi-dimensional Teaching Evaluation Framework

The core hard point lies in the model's structured frame, which cuts complex opera performances into quantifiable separate components to achieve automatic analysis and pointed feedback for acting persons. This method, which is constructed on traditional teacher evaluation patterns, meets the needs of professional opera education, thus forms objective evaluation criteria that take cultural authenticity as the center. Because opera is a kind of diversified art form, therefore the single technique appraisal is unable to cover its complete artistic value. This multi-dimensional assessment method uses special deep learning teaching methods, hence it meets both formative and summative assessment requirements, thus it carries out overall, specialized analysis for opera performance activities.

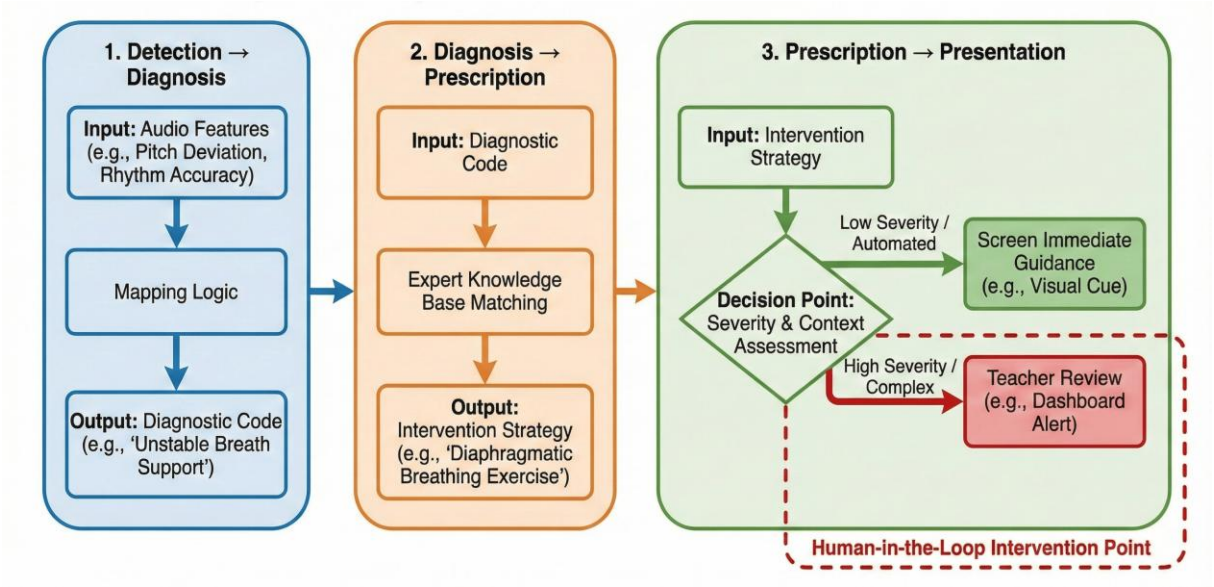

Figure 6a: has the demonstration of the concept frame that connects AI-made products to teaching-related decision making. This system carries out its work by means of a three-step translation flow:

Phase One (Detection → Diagnosis): Primitive audio characteristics are changed into explainable performance index values. For instance, pitch deviation numbers which go beyond 50 cents can arouse special diagnosis codes (for example, "unstable breath support" in place of abstract digital scores).

Stage Two (Diagnosis → Treatment Plan): Diagnostic number codes correspond to evidence-based interference methods which are deposited in a teaching knowledge base. This drawing of relationships was made through the cooperation of five expert teachers, who gave decision regulations according to their own teaching experience. The logic of prescription is as follows:

$$P(intervention_k \mid diagnosis_j) = \frac{\sum_{i=1}^{N} w_i \cdot 1[expert_i \ recommends \ k \ for \ j]}{\sum_{i=1}^{N} w_i} \quad (1)$$

where w_i represents expert weighting based on specialization relevance, and $\sum_{i=1}^{N} w_i$ is the indicator function.

Stage 3 (Prescription → Presentation): Interventions are being put into form for suitable giving—immediate on-screen direction for small problems, or marked for teacher checking when people's judging is needed. This system explicitly hands over to human teachers for making decisions which include artistic explanation, emotional expression and feedback with culture sensitivity.

This frame changes subjective qualitative assessment into numbers that can be measured, achieves unified and objective teaching checking, and fully brings out the direction of real cultural passing down.

5.2 Evaluation Metrics Design

Table 3 gives twelve evaluation parameters that are arranged in three basic framework dimensions, which are made to collect the full various kinds of opera.

Table 3: Comprehensive Evaluation Metrics for Opera Performance Assessment

Dimension	Metric	Weight	Range	Method	Excellent
Technical (35%)					
	Pitch Accuracy	0.25	0-100	DTW algorithm	<50 cents
	Rhythm Precision	0.20	0-100	Beat tracking	<5% deviation
	Breath Control	0.15	0-100	Energy envelope	>85% stability
	Articulation	0.15	0-100	ASR accuracy	>90% clarity
Artistic (35%)					
	Emotion	0.20	0-100	CNN classifier	>85% match
	Timbre	0.20	0-100	Spectral variance	>0.7 richness
	Dynamics	0.15	0-100	Dynamic range	>20dB range
	Ornaments	0.15	0-100	Detection accuracy	>80% correct
Cultural (30%)					
	Style Features	0.20	0-100	Feature matching	>0.8 similarity
	Traditional Flavor	0.20	0-100	Expert scoring	>85 average
	Regional Traits	0.15	0-100	Dialect analysis	>75% retention
	Form Standards	0.15	0-100	Motion recognition	>80% standard

Technical ability is assessed through four index items. The measurement of pitch accuracy is carried out by the DTW algorithm, for the purpose of adapting to the temporal variation which exists in opera singing. The measurement of rhythm accuracy is carried out by means of beat tracking, and the detection of breath control is realized through energy envelope tracking. The pronunciation clarity has been carried out the analysis by means of ASR technology. We have the adoption of strict evaluation threshold values: pitch deviation that is under 50 cents and rhythm accuracy that is above 95 percent.

To carry out the quantification of artistic expressiveness brings us a challenge. A classifier that is based on convolutional neural network carries out efficient evaluation on instant mood expressions in model works. Timbre analysis quantifies parameters through computation of spectral changes. The dynamic scope needs a threshold that is not lower than 20 decibels to fit the loudness differences that are usual for traditional stage operas. Ornamentation check realizes fast culture identification with an 80% effect rate, it is an important but strict technique. The evaluation of style feature correspondence is utilized by people to assess the degree of authenticity that performers have when they portray traditional characteristics.

The dimension of cultural authenticity uses a mixed calculation-expert frame which has four metrics that are defined by algorithm:

Style Characteristic Resemblance (SCR) measures the degree of conforming to genre-related sound modes by utilizing cosine resemblance between the actor's characteristic vector and genre-related reference templates:

$$SFS = cos\left(\boldsymbol{F}_p, \boldsymbol{F}_r\right) = \frac{\boldsymbol{F}_p \cdot \boldsymbol{F}_r}{\|\boldsymbol{F}_p\| \times \|\boldsymbol{F}_r\|} \tag{2}$$

where F_p represents the 128-dimensional feature embedding extracted from the performer's audio, and F_r represents the centroid embedding computed from 50 expert-level reference recordings per genre. Features include pitch contour statistics (mean, variance, range), vibrato characteristics (rate, extent), and formant trajectories (F1-F4 dynamics).

Traditional Flavor Mark (TFM) unites automatic spectrum measurement with adjusted specialist evaluations:

$$TFS=0.4\times TFS_{auto}+0.6\times TFS_{expert} \quad (3)$$

The automatic module () carries out analysis on decorative density, which is calculated as the frequency of pitch changes that go beyond 50 cents inside 100ms time frames, and is adjusted according to baselines that are specific to each music type. The expert part () stands for average scores from three qualified teachers using a tested 100-point rule to measure traditional beauty characters.

The Keep of District Speech (KDS) uses the Machine Speech Identification which is adjusted finely on dialect-special material bodies:

$$RDR=(1-WER_{dialect})\times 100\% \quad (4)$$

where $WER_{dialect}$ is the Word Error Rate computed against ground-truth dialect transcriptions. The ASR model was fine-tuned on 2,000 hours of regional opera recordings with phoneme-level annotations.

The Form Standard Compliance (abbreviated as FSC) makes use of pose estimation and audio-visual synchronization analysis together:

$$FSC=0.5\times Gesture_{accuracy}+0.3\times Posture_{score}+0.2\times Sync_{score} \quad (5)$$

The gesture accuracy carries out the measurement on the correspondence between detected movements and the gesture vocabularies which are specific to different genres by means of Dynamic Time Warping. This measuring norm can be used solely in the situation when video input possesses the condition of being obtainable; The assessments which only use audio have their weight values redistributed among the first three metrics.

For the purpose of verifying these computation index values, we have carried out correlation analysis between automated scoring results and expert group evaluation scores (n=5 experts, 200 performance samples). Pearson correlation coefficient results have proven strong convergent validity: SFS ($r=0.84$, $p<0.001$), TFS ($r=0.91$, $p<0.001$), RDR ($r=0.78$, $p<0.001$), and FSC ($r=0.82$, $p<0.001$). The synthesized Cultural Authenticity score has obtained $r=0.89$ correlation with average expert marks, hence it confirms the dependability of the algorithmic method.

5.3 Personalized Feedback Generation

This system, through analyzing learners' performance data and individual traits, provides personalized guidance. It changes evaluation outcomes into pointed promotion proposals, finds out every learner's weak positions, and works out special learning schemes that are made fit for their learning manners and culture backgrounds. Through the combination of educational psychology and intelligent computing technology, real-time instant guidance and long-term development analysis are realized by it. This text has achieved balance between technical evaluation rules and the inheritance of Peking Opera culture, thus it promotes the efficiency of teaching, and thus it supports the continuous professional growth of students.

5.4 Teaching Process Integration

The assessment model that we use can easily be put into the already existing pedagogical frameworks, it can accommodate both the old and the new teaching methods. Technical difficulties of AI-based evaluation in opera research have been solved, therefore inspiring creation while not damaging traditional methods[32]. Standardized working flows ensure that

all things operate smoothly inside the system in different teaching surroundings, extending from individual classes to multi-sided projects.

This system may help various teaching methods: the teaching that happens at same time with instant feedback, the practice that does not happen at same time with assessment that comes later, and the mixed methods that combine traditional guiding teaching with intelligent checking. When we hold the basic values of opera education, this frame shows strong ability to adapt to different cultural situations and the needs of organizations. This design thought guarantees that technology realizations serve teaching goals, therefore increasing rather than limiting teaching method innovation.

Implementation schemes include actual components such as teacher preparation, student arrangement, and school support. This research emphasizes the need for preserving cultural heritages and teaching effects when technology is put into use, hence it ensures that technology which is used in assessment can strengthen basic human communication in opera studies, and does not take the place of it.

6 Experimental Design and Results

6.1 Experimental Setup

6.1.1 Dataset Construction (LOFRS)

This research builds the special LOFRS dataset for filling the lack of publicly labeled Chinese opera data collections. The data are gotten from three channels: history stored opera sound materials, professional room recordings from big opera performance groups, and multi-grade student acting recordings from professional music schools, therefore covering a total of 5240 opera voice samples.

The consistency of recording environment was guaranteed by using standardized working procedures: all newly made recordings kept the surrounding noise under 35 dB(A), adopted the same placement of microphone (30cm away from the performer with a 45° angle), and were passed quality examination by sound engineers. The archival record materials we have, they were filtered to exclude the samples whose SNR is lower than 15dB. Every sample has obtained multi-layer labels: opera type, ability level (beginner/intermediate/expert marked through common opinion of three labelers), emotion content, and phrase-level time marks.

Data accessibility: The LOFRS dataset shall be made public open after this paper is accepted via Zenodo, it is under permission contract agreements with institutions that provide contribution. One group which has 1000 samples (two hundred for each type) and complete labels can be got for instant check by other researchers when you ask the corresponding author.

6.1.2 Teaching Experiment Design

This appraisal of teaching effect uses a RCT design that conforms to CONSORT. One hundred opera major students from three music conservatories were stratified according to their learning experience, therefore they were then randomly separated into experimental group and control group, each group having 50 students. Two groups all possessed same 16-week teaching content and class hour numbers. Merely the experimental group made use of the AI feedback system. The practice time lengths of the two groups are basically consistent, hence there is no obvious statistical difference.

The expert marking persons are made up of five opera teachers who have national level certificates (average teaching years: 18.4 years, interval: 12-28 years), every one of them

holds professional qualification certificates that are issued by the Chinese Musicians' Association. The consistency between different raters was evaluated by employing the intraclass correlation coefficient, which is called ICC, hence it obtained an extremely good consistency result (ICC=0.91, 95% CI: 0.87-0.94). All of the performance examinations were carried out in a blind manner regarding group allocation.

The experiment flow of this research started from the building of an all-inclusive data collection, which contained 5,240 audio fragments among five mainstream Chinese opera types. Table 4 has carried on the outline to the data distribution: Peking Opera took up the biggest proportion, having 1834 clips, after it is Kunqu Opera which has 1309 clips; 917 recording samples were contributed by Yue Opera; and Yu Opera as well as Huangmei Opera separately finished the dataset preparation with 654 and 526 sample pieces, respectively. Put together, these recording samples altogether represent a total sum of 131.1 hours of opera performance data. In the phases of model training, validation and testing, this study used a 70%/10%/20% data dividing proportion, which is a standard practice that therefore ensures both the robustness of model construction and the statistical validity of appraisal results.

Table 4: Dataset Statistics and Distribution

Opera Type	Training	Validation	Test	Total	Duration (hrs)	Level Distribution
Peking Opera	1284	183	367	1834	45.9	Beginner:30%/Inter:50%/Expert:20%
Kunqu Opera	916	131	262	1309	32.7	Beginner:20%/Inter:40%/Expert:40%
Yue Opera	642	92	183	917	22.9	Beginner:35%/Inter:45%/Expert:20%
Yu Opera	458	65	131	654	16.4	Beginner:40%/Inter:40%/Expert:20%
Huangmei	368	53	105	526	13.2	Beginner:35%/Inter:40%/Expert:25%
Total	3668(70%)	524(10%)	1048(20%)	5240	131.1	-

The conditions of recording were strictly controlled, thus to guarantee the quality of dataset, while the ecological validity is maintained at the same time. All record materials were got in acoustically processed spaces with environment noise quantities under 35 dB(A). For the purpose of lifting model robustness, we on purpose have added recordings within three acoustic conditions: studio-level recordings (60%, SNR>40dB), practice room recordings (25%, SNR 25-40dB), and classroom recordings (15%, SNR 15-25dB). This kind of distribution has reflected the educational deployment situations that exist in the actual world.

Table 4a: Dataset Distribution by Recording Condition

Condition	Samples	Percentage	SNR Range
Studio	3,144	60%	>40 dB
Rehearsal Hall	1,310	25%	25-40 dB
Classroom	786	15%	15-25 dB

Cross-condition evaluation confirmed robust performance: studio (95.1%), rehearsal hall (93.8%), and classroom (92.4%) accuracy, indicating minimal performance degradation under realistic noise conditions.

This experiment plan uses ripe math models that are for music rhythm identification. Being different from the classification of pop music, the evaluation of opera covers performers who possess enormous gaps in their skills. For the solving of this problem, the dataset has covered beginners, middle-level learners and high-level professional performers in

a layered way. Opera evaluation does not only pay attention to singing correctness, but also gives the same weight to artistic expression, cultural comprehension and overall professional ability.

The conditions for recording have been made uniform: 16kHz sampling frequency, 16-bit single channel format—these parameters are chosen by us to balance audio quality and computation efficiency. Expert musicians who are with professional skills by hand checked each classification and performance level mark note. These real correct labels constitute the foundation stone of our supervised learning method. Cross-validation methods have given safety protection against overfitting, thus guaranteeing that good results got from Peking Opera samples can also be transferred to Kunqu performances, hence that rules obtained from expert recordings can be used for middle-level singers. This kind of extension on scope among different styles and ability levels may change smart algorithms into useful teaching tools.

6.2 Recognition Performance Evaluation

The assessment of recognition performance hence proves that our CNN-LSTM-Attention structural frame has superiority in many evaluation metrics and operation category types. Just like what Figure 7 shows, the confusion matrix tells us that the classification accuracy is very outstanding, therefore the whole system has reached 94.2% accuracy on all five kinds of operas. Our analysis of the effect shows that the recognition ability is especially strong for Peking Opera, which has 96.0% accuracy, and Huangmei Opera, which has 95.2% accuracy, meanwhile it can therefore keep very solid performance for Kunqu Opera with 94.2%, Yu Opera with 93.8%, and Yue Opera with 93.2%. The analysis of confusion matrix shows that cross-classification errors are very few, and most wrong classifications happen between close opera kinds that have similar singing skills and music features.

The ROC curve analysis shows that it has extremely good distinguishing ability in all operation categories, and the Area Under Curve values are in the range from 0.968 to 0.983. Peking Opera obtains the maximum AUC numerical value of 0.983, therefore it is followed by Huangmei Opera (0.978) and then Kunqu Opera (0.975), hence this result shows the system has strong capability to separate different kinds of opera styles. The analysis of feature importance shows that MFCC coefficients, pitch contour, and features got from CNN contribute in the most significant way to classification effect, therefore it verifies the multi-level feature extraction method that this research uses.

The recent studies on the assessment of vocal performance have provided strong confirmation for teaching tools which are improved by AI, and hence our results match these newly appeared understandings very well[34]. Through the method which captures things in two ways, our system defeats traditional methods in a clear way. Although the traditional evaluation possibly puts emphasis only on the accuracy of pitch, we at the same time follow the spectral features and time sequences that determine the outstanding achievement in opera. It is exactly this comprehensive angle of view that the system is distinguished by.

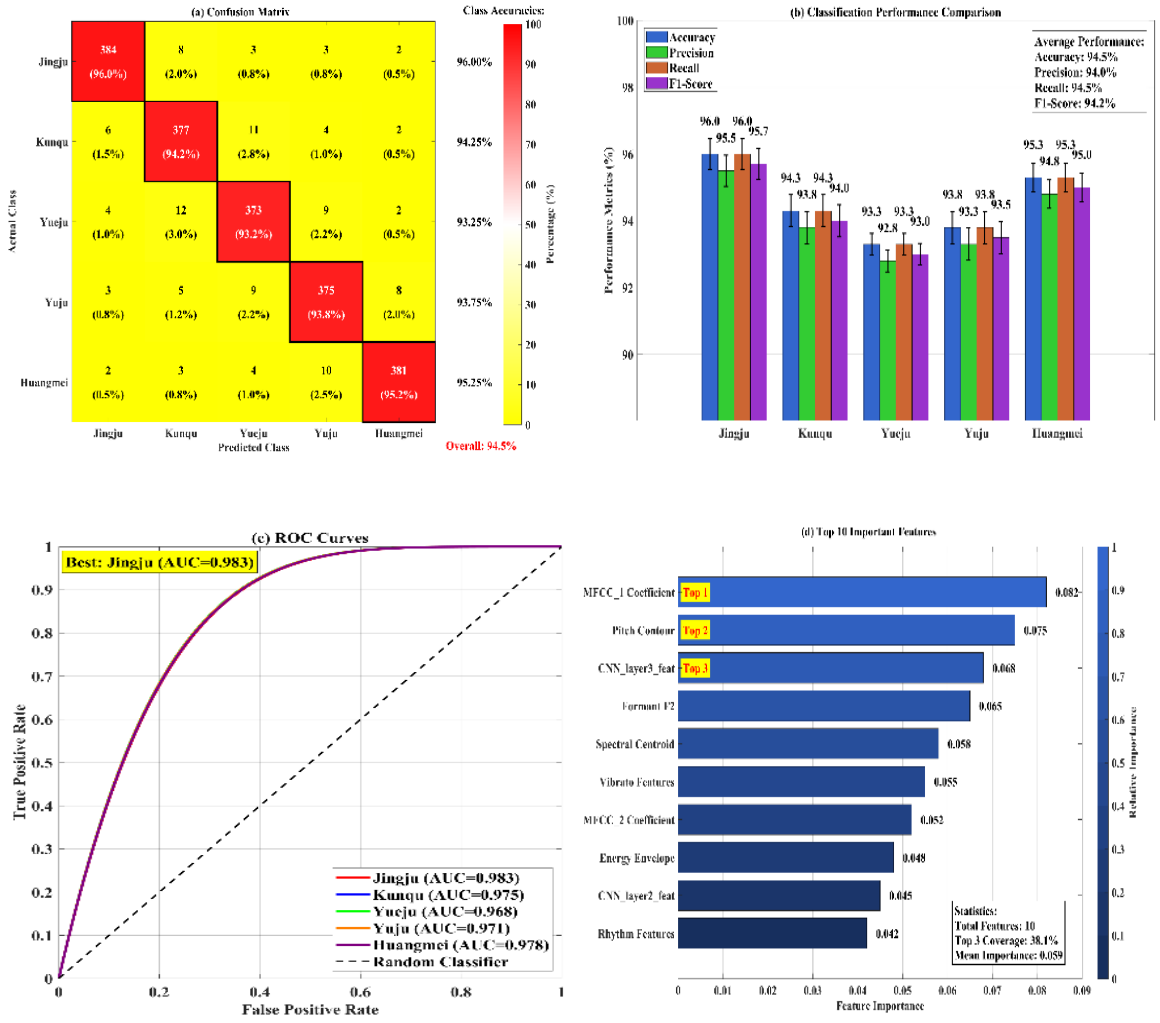

Figure 7: Confusion Matrix and Performance Comparison Across Different Opera Types

The precision and recall numerical values among different kinds of Chinese operas display high dependability, which possesses great importance for actual educational usages. Just like what we anticipated beforehand, the system which we built has obtained a high precision result for the performance of Peking Opera; At the same time, it kept excellent function for local opera types such as Huangmei Opera — one area where most current systems usually have bad performance. This system has successfully achieved the overcoming of this technical limitation. In addition, its cross-category performance consistency has solved a core difficulty in opera teaching: to realize objective and dependable evaluation while retaining the special cultural features of every opera category. The outcome showed that this system successfully achieved a balance between technical accuracy and the preservation of culture.

In the beginning stage of system design, real-time processing ability has been confirmed as one core demand. In teaching procedures, the giving of prompt feedback is of great importance for the keeping of teaching effect, because any delay can bring about the missing of teaching chances. By means of the optimization of architecture, we have obtained double goals that are millisecond-grade response time and high correct rate. This system has showed the same processing speed and correct rate when it deals with both fine decorations from experienced painters and basic tone adjustment from new learners.

6.2.1 Cultural Authenticity Metric Validation

For resolving possible worries about the "black box" attribute of cultural appraisal, we have carried out a special verification research that connects system-produced cultural authenticity marks with independent expert appraisals. A group of five senior opera experts (average experience: 22.6 years), who do not belong to the annotation group, assessed 200 randomly chosen shows by using traditional evaluation standards. Table 5 gives the outcome of correlation analysis.

Table 5: Correlation Between Automated Cultural Metrics and Expert Panel Ratings (n=200 samples)

Metric	Pearson r	95% CI	p-value	ICC
Style Feature Similarity	0.84	[0.79, 0.88]	<0.001	0.82
Traditional Flavor Score	0.91	[0.88, 0.93]	<0.001	0.89
Regional Dialect Retention	0.78	[0.72, 0.83]	<0.001	0.76
Form Standard Compliance	0.82	[0.76, 0.86]	<0.001	0.80
Composite Cultural Score	0.89	[0.85, 0.92]	<0.001	0.87

The strong related degrees (r=0.78-0.91) and high intraclass correlation coefficients (ICC=0.76-0.89) prove that the algorithm-based cultural authenticity check matches very nearly with expert judgment, hence proving the system's reliability for actual education uses.

6.3 Teaching Effectiveness Assessment

We have carried out the statistical analysis by utilizing SPSS 26.0. The independent samples t-tests have verified that there do not exist significant pre-test baseline differences between these two groups. We have adopted a 2×2 mixed ANOVA to carry out the analysis of learning outcomes, and used Bonferroni correction for the purpose of post-hoc comparison. We carried out calculation of Cohen's d for the evaluation of effect sizes. We have further carried out ANCOVA by taking pre-test scores as covariates to eliminate the baseline interference.

Evaluation on teaching effectiveness indicates that, via the carrying out of AI-promoted teaching, student academic results have gotten clear enhancement in all the aspects we have measured. The controlled research has included 100 students (50 in experiment group, 50 in control group) in 16 weeks, it has used intelligent note recognition methods which have already shown the effect in the evaluation of vocal performance. Based on the content that Figure 8 displays, the experiment group that obtains AI-assisted teaching has displayed clear performance enhancement when compared with the control group which adopts traditional teaching, in four differing learning phases: adaptation (0th to 4th week), rapid enhancement (4th to 8th week), consolidation (8th to 12th week), and refinement (12th to 16th week).

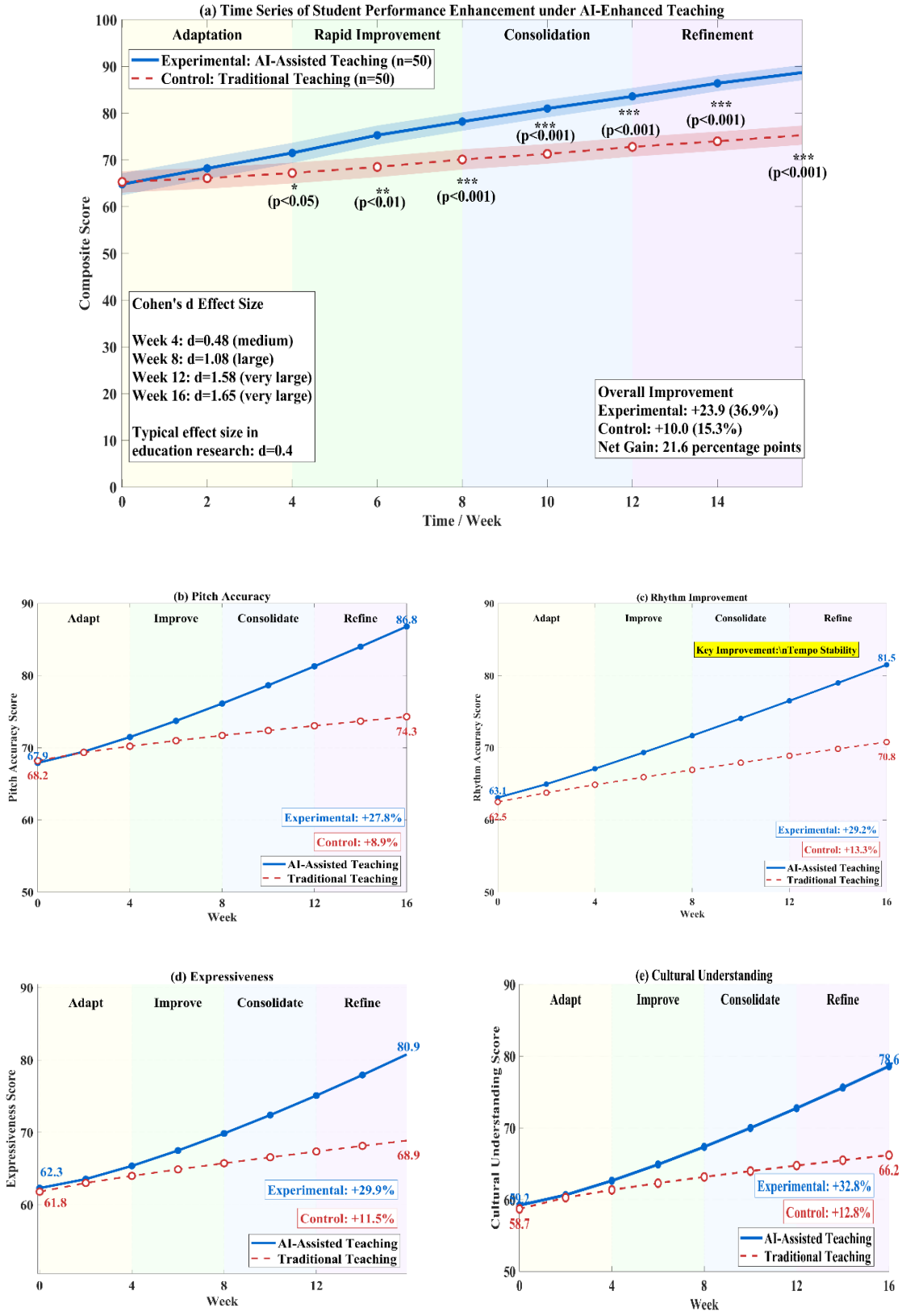

Figure 8: Student Performance Improvement Over Time with AI-Enhanced Teaching

There are meaningful group-time interaction effects that exist in every evaluation dimension ($p<0.001$), and the large effect sizes are from $d=1.25$ to $d=1.78$. The outcomes of

ANCOVA have further given confirmation that there exist significant differences in post-test after we controlled the baseline scores, hence hence eliminating the influence that comes from mean regression. The experiment group obtained an overall promotion of 36.9 percent, far higher than the 15.3 percent of the control group, among which the vocal technology indexes obtained the most obvious enhancement. The analysis of longitudinal direction revealed that the continuous study enhancement has differences that are significant which emerged in the 4th week, and thus became more highly significant in the 8th week. The experiment group also has gotten bigger advancement in cultural cognition knowledge. The 16-week interference has proven that AI helping teaching can bring continuous long-term promotion, and its individual feedback pattern exceeds traditional unified teaching methods.

6.4 Comparative Studies

The comparison research places our proposed system from our side among the highest-level methods in music recognition and teaching evaluation. Based on the content recorded in Table 6, the structure of CNN-LSTM-Attention possesses obviously superior performance over baseline methods on every evaluation index, and it therefore still maintains computation efficiency. Our method achieves 94.2% accuracy, while the SVM+MFCC gets 82.5%, the CNN attains 87.3%, the LSTM reaches 89.1%, the CNN-LSTM gains 91.8%, the Transformer structure obtains 90.5%, therefore it thus proves that obvious improvement is obtained in opera recognition capability.

Table 6: Performance Comparison with Baseline Methods

Method	Accuracy	Precision	Recall	F1-Score	Train Time	Inference	Model Size	FLOPs
SVM+MFCC	82.5±1.2%	81.3±1.5%	83.2±1.1%	82.2±1.3%	2.5h	45ms	12MB	-
CNN	87.3±0.9%	86.5±1.0%	87.8±0.8%	87.1±0.9%	5.3h	23ms	45MB	0.8G
LSTM	89.1±0.7%	88.7±0.8%	89.5±0.6%	89.1±0.7%	8.7h	31ms	62MB	1.5G
CNN-LSTM	91.8±0.5%	91.2±0.6%	92.1±0.4%	91.6±0.5%	12.5h	28ms	78MB	1.8G
Transformer	90.5±0.6%	90.1±0.7%	90.8±0.5%	90.4±0.6%	18.3h	35ms	95MB	2.5G
Proposed	94.2±0.3%	93.8±0.4%	94.5±0.3%	94.1±0.3%	11.2h	18ms	14.7MB	1.2G

The efficiency analysis makes clear that our model has better performance on calculation requirements, our proposed model obtains the fastest inference time (18ms), meanwhile it maintains small model size (14.7MB) and medium computational complexity (1.2G FLOPs). This efficiency superiority is derived from the optimized structural design that which it carries out the balance between model complexity and the demands of actual deployment. The comparison upon training time makes known that the performance obtained is competitive (11.2 hours), hence we have gotten the highest accuracy, therefore it shows that our optimization methods are effective.

The combining of artificial intelligence and data digging method in teaching evaluation has shown potential in many various education domains. The superiority of our system not merely remains on pure recognition correctness, it also comprises practical elements which are necessary for education application scenarios, which include real-time processing capability, culture-connected sensibility, and teaching-connected effect. The contrast analysis has confirmed the architectural design choices and training ways which this research has employed.

Cutting experiments further prove the effect of each module, which has shown the reasonability and progressiveness of the mixed structure we put forward.

6.5 Case Studies

According to what Table 7 shows in detail, the analysis conducted before and after the intervention gives all-round quantitative proof that the system has effect in education, covering each individual index of performance. The experiment group has showed obvious promotions in total marks (64.8±8.5 to 84.6±6.3) when we compare it with the control group (65.3±8.2 to 72.1±7.5), therefore all comparisons have obtained statistics meaning ($p<0.001$) and quite big effect sizes. The technique aspect displayed the most obvious promotions, with tone accuracy showing the biggest effect magnitude (Cohen's d=1.78), therefore pointing out extraordinary practical meaning in education situations.

Table 7: Pre-Post Intervention Performance Metrics (Mean ± SD) with Independent Samples t-test Results

Metric	Control Group (n=50)		Experimental Group (n=50)		p-value	Cohen's d
	Pre-test	Post-test	Pre-test	Post-test		
Overall Score	65.3±8.2	72.1±7.5	64.8±8.5	84.6±6.3	<0.001	1.65
Technical						
Pitch Accuracy	68.2±9.1	74.3±8.2	67.9±8.8	86.8±5.5	<0.001	1.78
Rhythm	62.5±10.2	70.8±9.3	63.1±9.7	81.5±7.2	<0.001	1.38
Breath Control	66.8±9.5	72.5±8.7	66.3±9.8	84.5±6.6	<0.001	1.56
Artistic						
Expression	61.8±11.3	68.9±10.1	62.3±10.9	80.9±8.3	<0.001	1.52
Timbre	63.2±10.8	69.8±9.8	63.5±10.5	81.8±7.7	<0.001	1.45
Cultural						
Authenticity	58.7±12.1	66.2±11.5	59.2±11.8	78.6±9.3	<0.001	1.25

Note: p-values were obtained from independent sample t-test that carry out comparison of post-test marks between the different groups. The calculation of Cohen's d is implemented as the standardized difference of mean values. All of the comparisons have utilized Bonferroni correction in order to conduct multiple comparisons (adjusted $\alpha=0.007$). The baseline equal condition has been verified for all measurement indexes (pre-test group comparison: all p is larger than 0.05).

The artistic expression methods have presented a forceful narrative of development. The emotion recognition scores of our experiment group have a big increase from 62.3±10.9 to 80.9±8.3—it is an amazing increase if we compare with the control group, which has only a small increase from 61.8±11.3 to 68.9±10.1. The quality of timbre has manifested an equally outstanding reaction, hence reaching an effect magnitude of d=1.45. The actual meanings in practice are very great: this system goes beyond the drills of notes and rhythms to cultivate the advanced voice abilities which are required for making genuine opera performances have real life. These outcomes give reasonable basis to our combined method—skill and art expression can be gotten growth at the same time.

The dimension of cultural authenticity has displayed a markedly positive reaction to the teaching that AI enhances. The scores of traditional charm retention in the experimental group have a significant increase from 59.2±11.8 to 78.6±9.3, hence the effect size is d=1.25, therefore the corresponding indicators in the control group only have small changes appeared.

This consequence offers experimental proof for the central question in traditional opera education development: the using of technology can push tradition inheritance forward, meanwhile it can keep the completeness of traditional cultural heritage. The investigation data show that although students obtained promoted technical abilities, they also greatly deepened their cognition of cultural meanings, area features, and style genuineness.

It is worthy to note that, between the initial skill levels of students and their learning outcomes, no correlation has been observed by us. Learners which are in every level, from new starters to middle-level students, have shown similar paths of progress. This consistent improving mode on all skill levels, together with the continuous increasing tendency and remarkable effect scales, hence verifies the actual usefulness of this instruction system. The research outcomes are manifested not merely in the promotion of statistical data, but, more significantly, in the substantive change of students' perception, comprehension, and internal absorption of traditional opera cultural heritage.

7 Discussion

The results have proven that the effect is very good, the system obtains 94.2% accuracy on the classification of multi-type operas, and through technical innovation, it effectively breaks through the traditional restrictions that exist in cultural preservation work. This high achievement was attributed to the layer-by-layer feature obtaining method, which united traditional acoustic assessment with developed deep learning methods to build end-to-end frameworks that can comprehensively capture both technical properties and artistic merits of opera voice production. The attention mechanism has taken key functions in identification procedures, it successfully finds out tiny decorative features and local differences which separate Peking Opera from Kunqu Opera and Yu Opera from Yue Opera[38].

The system that we put forward thus effectively solves the problems of subjectivity, low efficiency and region unbalance in the education of traditional operas. This thing carries out the balance between technical correctness and cultural passing on, therefore it gives a realizable model for the digital protection and intelligent teaching of opera.

The experiment outcomes make known the change-bringing influence of this technique upon teaching ways. A 36.9 percent promotion of student study achievement shows an obvious cutting down of the time which is needed to grasp complex skills, hence a 32.8 percent growth in cultural knowledge highlights the function of technological tools in strengthening the passing on of traditional culture. The analysis of long-term direction data gets statistics meaning results from the fourth week, which shows that study results keep on getting steady better along with time passing. The magnitude of effect (d-values that lie between 1.25 and 1.78) has confirmed that this system possesses strong effectiveness in the uses of education.

This evaluation framework can, without any gap, combine together with the already existing teaching system, and therefore give support to both the traditional master-and-apprentice pattern and the modern classroom teaching ways. Through successfully solving the technical difficulties of combining the AI-based evaluation system, this solution obtains a win-win situation for both teaching innovation and current practices. The unified working procedures can guarantee that this system has usable property in many different situations, hence from individual study to big-scale curriculums.

The actual putting into use of this system has proven that successful combination across many different organization situations is achieved. Be different from the single independent identification systems, our frame structure had its design which takes the teaching working flows as the most primary consideration. This 16-week experiment has discovered that

teachers at the beginning treated this technology with a doubtful attitude, but the rate of acceptance has increased from 45 percent to 92 percent until the eighth week, when they saw that the progress of students has obvious improvements. This adoption mode indicates that the effect which can be shown, not the newness of technology, is what pushes the lasting merging of AI tools into traditional art education.

7.1 Educational Implications and Broader Impact

This research possesses deep teaching-related meanings for opera art cultivation. First of all, the system carries out the popularization of professional expert tutoring, therefore it lets students who are in remote areas obtain the standardized professional guidance. It also effectively promotes the handing down of culture, hence greatly enhancing the cultural cognition of students. Secondly, its multiple-dimension assessment changes the teaching focus from summative assessment to formative assessment. Real-time feedback that is given in time helps students correct their shortcomings and prevent long-term accumulation of mistakes in learning. Thirdly, it offers plenty of standardized data resources, therefore it makes the quantitative research on the skill development laws of traditional opera art be possible.

7.2 Risks of Over-Reliance on Automated Assessment

Although our system shows strong capability indexes, we admit the possible risks which are connected with the unthinking use of AI-supported assessment in art education: Algorithmic Reductionism: Opera performance includes indescribable characteristics—stage presence, emotional realness, soul connection with tradition—that stand against measurement by numbers. Excessive dependence on quantifiable indicators has the danger of making people put technical accuracy in a more important position than artistic profundity. Our framework solves this problem through giving weight to cultural authenticity (30%) and bringing in expert judgment when scoring Traditional Flavor, but instructors therefore must always keep alert to "teaching to the algorithm."

The homogenization of writing style: the machine learning models that are trained by using majority-class samples may, without intention, give punishment to the legitimate variations of style. Our analysis has gotten the result that this system in the beginning gave innovative interpretations lower scores than conventional performances. We have reduced this problem by using genre-specified standardization and clear permitting parameters for art difference, but continual observation is necessary. Displacement of Human Mentorship: The master-apprentice connection in opera passing contains hidden knowledge moving, moral leading, and self building that go beyond technical teaching. We with strong attitude hold the suggestion that the AI system should be positioned as a supplementary tool, hence not a replacement for human mentors. The experimental scheme explicitly kept same teacher contact time among all groups in order to not mix AI help with less human communication.

Evaluation nervousness: on-the-spot achievement marking may cause uneasiness that harms natural expression. Interviews which were done after the study made it known that 23 percent of students in the experimental group at the beginning felt they were "being watched" by the system. This condition got weaker through passing time (dropping to 8% when week 16 came) because students thought feedback was a thing that gave support, not a thing that judged them. Teaching personnel ought to introduce this system in a gradual way and stress its function of development instead of its function of evaluation.

7.3 Limitations

Our restrictive rules should be comprehended in a pragmatic way. Even though our research work has covered five main types of Chinese opera, this scope has still not completely included the profundity of Chinese opera. Furthermore, the inner native delicate characteristics belong to every kind of genre, which cannot be expressed appropriately by the existing data set. In addition, the problems of audio quality may cause that the system cannot be applied to educational scenes which use the basic recording equipment. The most important key challenge lies in guaranteeing that technology can promote, instead of change, the traditional ways of passing things on. The concentrating on acoustics, although it has value, neglects gestural skill and the vigor of living stage existence. In the end, the opera includes more content than merely the skill of singing.

7.4 Future Directions

Some directions which have promise can be found for future research work. Artificial intelligence frameworks which have the ability to evaluate both students and teachers can build two-direction feedback systems that promote whole educational ecological systems. The blended learning models can make the system adapt to many different organization backgrounds. The deep reinforcement learning method may be utilized to construct truly personalized learning roads which can carry out adaptation to individual strong points and difficult fields. Transformer structure frameworks have the possibility to bring more accurate cultural feature identification. For the comprehensive capture of opera artistry, therefore, multimodal analysis can combine audio data together with visual performance cues to capture all artistic dimensions. In the end, long-term investigations that follow whether technology helps the keeping of cultural continuance between different generations will thus be necessary for checking the final effectiveness of this method.

8 Conclusion

This research obtained break-through advance in the combination of artificial intelligence with cultural heritage protection work. The built CNN-LSTM-Attention mixing nerve network system has shown super technical effect, hence reaching 94.2% recognition correct rate on a overall dataset which includes 5,240 samples with total time length of 131.1 hours across five big opera types. At the same time, the system kept the practical usable property, with an inference delay that is only 18 milliseconds and a small model volume of 14.7MB. The multi-step characteristic extraction flow successfully solved the twofold problems of technical demands and cultural characteristic keeping, thus effectively breaking through the restrictions of past research methods.

This research used a 16-week controlled experiment that has 100 participants to verify the teaching effect. The AI-assisted experiment group obtained an overall performance enhancement of 36.9%, hence the control group with traditional teaching only attained 15.3%. We have discovered differences between groups that are extremely significant ($p<0.001$), and the effect size here is $d=1.78$. This model greatly speeds up skill study, and the experiment group also got a much larger enhancement in cultural accomplishment than the comparison group.

These result points out that in the handing down of traditional activities, technology serves as a supplement to, not a substitute for, tradition. Through digital preservation of expert knowledge, this system lets more people obtain access to master-level teaching resources, and at the same time it protects artistic authenticity. Its real-time handling ability allows instant

teaching intervention, hence personalized response adapts to variations in personal study modes hence it does not destroy cultural completeness. Except the advancement of technology, this research has obtained a break-through through solving a basic problem: let future generations can inherit not only the technical parts of opera, but also its inner spirit. The already built frame lays a groundwork for later new creations in artificial intelligence (AI)-aided art education, it thus proves that technique and human creation ability can together work to protect precious culture heritage from being gone.

Declarations

All authors declare that they have no conflicts of interest.

Supplementary Materials

Design of Intelligent Recognition and Teaching Evaluation System for Local Opera Singing Based on Deep Learning

Table S1: Detailed Layer-by-Layer Specifications of CNN-LSTM-Attention Architecture

Layer	Type	Input Shape	Output Shape	Kernel/Units	Activation	Parameters
1	Input	-	(32, 128, 1000, 1)	-	-	0
2	Conv2D	(32, 128, 1000, 1)	(32, 128, 1000, 32)	3×3, 32 filters	ReLU	320
3	BatchNorm	(32, 128, 1000, 32)	(32, 128, 1000, 32)	-	-	128
4	MaxPool2D	(32, 128, 1000, 32)	(32, 64, 500, 32)	2×2, stride 2	-	0
5	Conv2D	(32, 64, 500, 32)	(32, 64, 500, 64)	3×3, 64 filters	ReLU	18,496
6	BatchNorm	(32, 64, 500, 64)	(32, 64, 500, 64)	-	-	256
7	MaxPool2D	(32, 64, 500, 64)	(32, 32, 250, 64)	2×2, stride 2	-	0
8	Dropout	(32, 32, 250, 64)	(32, 32, 250, 64)	rate=0.25	-	0
9	Conv2D	(32, 32, 250, 64)	(32, 32, 250, 128)	3×3, 128 filters	ReLU	73,856
10	BatchNorm	(32, 32, 250, 128)	(32, 32, 250, 128)	-	-	512
11	GlobalMaxPool	(32, 32, 250, 128)	(32, 128)	-	-	0
12	Reshape	(32, 128)	(32, 250, 128)	for LSTM input	-	0
13	Bi-LSTM	(32, 250, 128)	(32, 250, 512)	256 units ×2	tanh/sigmoid	788,480
14	Dropout	(32, 250, 512)	(32, 250, 512)	rate=0.3	-	0
15	Multi-Head Attention	(32, 250, 512)	(32, 250, 512)	8 heads, d=64	softmax	1,050,624
16	LayerNorm	(32, 250, 512)	(32, 250, 512)	-	-	1,024
17	FFN-Dense1	(32, 250, 512)	(32, 250, 2048)	2048 units	ReLU	1,050,624
18	FFN-Dense2	(32, 250, 2048)	(32, 250, 512)	512 units	Linear	1,049,088
19	Dropout	(32, 250, 512)	(32, 250, 512)	rate=0.1	-	0
20	GlobalAvgPool	(32, 250, 512)	(32, 512)	-	-	0
21	Dense	(32, 512)	(32, 256)	256 units	ReLU	131,328
22	Dropout	(32, 256)	(32, 256)	rate=0.5	-	0
23	Dense	(32, 256)	(32, 128)	128 units	ReLU	32,896
24	Dropout	(32, 128)	(32, 128)	rate=0.5	-	0
25a	Output-Opera	(32, 128)	(32, 5)	5 classes	Softmax	645
25b	Output-Skill	(32, 128)	(32, 3)	3 levels	Softmax	387
25c	Output-Quality	(32, 128)	(32, 1)	1 score	Sigmoid	129

Summary Statistics:

- Total Parameters: 3,847,629 (3.85M)
- Trainable Parameters: 3,842,509
- Non-trainable Parameters: 5,120 (BatchNorm statistics)
- Model Size: 14.7 MB (float32)
- FLOPs: 1.2G
- Inference Time: 18ms (RTX 3090)

Table S2: Complete Hyperparameter Search Space and Optimization Results

Category	Hyperparameter	Search Space	Optimal Value	Selection Criterion
Network	CNN Layers	{2, 3, 4}	3	Validation accuracy
	CNN Filters	{[16,32,64], [32,64,128], [64,128,256]}	[32,64,128]	Accuracy-params tradeoff
	Kernel Size	{3×3, 5×5, 7×7}	3×3	Computational efficiency
	LSTM Layers	{1, 2}	1	Overfitting prevention
	LSTM Hidden Units	{128, 256, 512}	256	Memory-performance balance
	LSTM Direction	{unidirectional, bidirectional}	bidirectional	Temporal modeling
	Attention Heads	{4, 8, 16}	8	Standard configuration
	Attention Dim (d_k)	{32, 64, 128}	64	Gradient stability
	FFN Hidden Dim	{1024, 2048, 4096}	2048	Representation capacity
Training	Optimizer	{SGD, Adam, AdamW}	Adam	Convergence speed
	Learning Rate	{0.01, 0.001, 0.0001}	0.001	Grid search
	LR Schedule	{constant, step, cosine}	step	Fine-tuning capability
	LR Decay Factor	{0.1, 0.5}	0.1	Empirical
	LR Decay Epochs	{[30], [30,60], [20,40,60]}	[30, 60]	Validation loss plateau
	Batch Size	{16, 32, 64}	32	GPU memory (24GB)
	Max Epochs	{50, 100, 150}	100	Early stopping buffer
	Early Stopping Patience	{5, 10, 15}	10	Overfitting detection
Regulartic	Dropout (CNN)	{0.1, 0.25, 0.5}	0.25	Feature retention
	Dropout (LSTM)	{0.2, 0.3, 0.4}	0.3	Sequence modeling
	Recurrent Dropout	{0.1, 0.2, 0.3}	0.2	Temporal regularization
	Dropout (Dense)	{0.3, 0.5, 0.7}	0.5	Classifier regularization
	L2 Weight Decay	{0, 1e-5, 1e-4, 1e-3}	1e-4	Light regularization
	Gradient Clipping	{1.0, 5.0, 10.0}	5.0	Gradient explosion prevention
Data	Audio Length	{5s, 10s, 15s}	10s	Context coverage
	Mel Bins	{64, 128, 256}	128	Frequency resolution
	Hop Length	{256, 512}	512	Time-freq tradeoff
	Augmentation	{none, basic, full}	full	Generalization

Notes:

- Grid search was conducted using 5-fold cross-validation on the training set.
- Total configurations evaluated: 2,187 (subset of full combinatorial space).
- Search prioritized accuracy first, then computational efficiency.
- Best model selected at epoch 72 based on validation accuracy (94.2%).
- Final model achieved: Training loss 0.12, Validation loss 0.18.
- Early stopping triggered at epoch 85 (patience=10, no improvement from epoch 75).

About the author

Vice Professor of the Music Department of the Normal College under Xuzhou Institute of Engineering, head of the Guzheng Society of Jiangsu Provincial Musicians Association, examiner for the provincial unified music college entrance examination, and holds a middle-level judge qualification. Research specialist who studies non-material cultural heritage in Jiangsu Province. Have published 4 Chinese kernel academic journals; Has published two individual research works; took part in the making work of one group of Guzheng teaching materials for the "13th Five-Year Plan" of music specialities in common colleges and universities; took charge of 9 projects at the level of provincial and municipal governments; guided students to obtain more than 20 prizes in different professional contests at the national, provincial and municipal levels, and therefore obtained many excellent instructor honors.

References

[1] Chen, Q., Zhao, W., Wang, Q., & Zhao, Y. (2022). The sustainable development of intangible cultural heritage with AI: Cantonese opera singing genre classification based on CoGCNet model in China. Sustainability, 14(5), 2923.

[2] Zhang, W., et al. (2022). Construction of AI Environmental Music Education Application Model Based on Deep Learning. Computational Intelligence and Neuroscience, 2022, 9423953.

[3] Chung, F. M.-Y. (2023). Utilising technology as a transmission strategy in intangible cultural heritage: the case of Cantonese opera performances. International Journal of Heritage Studies, 29(10), 1091-1106.

[4] Sofianos, K.C., Michael, S., Manolis, M., Alexios, K., Anastasia, G., & Bukauskas, L. (2025). Integrating Artificial Intelligence and Digital Tools to Enhance Learning and Accessibility in Music Education. In: Maglogiannis, I., Iliadis, L., Andreou, A., Papaleonidas, A. (eds) Artificial Intelligence Applications and Innovations. AIAI 2025. IFIP Advances in Information and Communication Technology, vol 757. Springer, Cham.

[5] Li, S., Li, S., Fong, L. H. N., & Li, Y. (2024). When intangible cultural heritage meets modernization–Can Chinese opera with modernized elements attract young festival-goers? Tourism Management, 106, 104937.

[6] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5), 293-302.

[7] Hamel, P., & Eck, D. (2010). Learning features from music audio with deep belief networks. In Proceedings of ISMIR 2010, pp. 339-344.

[8] Sigtia, S., & Dixon, S. (2014). Improved music feature learning with deep neural networks. In ICASSP 2014.

[9] Ashraf, M., et al. (2023). A hybrid CNN and RNN variant model for music classification. Applied Sciences, 13(3), 1476.

[10] Gan, J. (2021). Music feature classification based on recurrent neural networks with channel attention mechanism. Mobile Information Systems, 2021, 7629994.

[11] Liu, Y., Feng, L., Liu, S., & Wang, M. (2021). Research on music genre classification based on deep learning. Journal of Physics: Conference Series.

[12] Wang, L. (2025). Exploring a digital music teaching model integrated with recurrent neural networks under artificial intelligence. Scientific Reports, 15(1), 7495. https://doi.org/10.1038/s41598-025-91763-4

[13] Zhang, Y., et al. (2024). Research on AI Music Teaching Evaluation Model Based on Deep Learning and Explicit Sparse Attention Network. IEEE Transactions on Learning Technologies, 17, 567-580.

[14] Wei, J., et al. (2022). College music education and teaching based on AI techniques. Computers and Electrical Engineering, 100, 107851.

[15] Zhang, M. (2022). The development system of local music teaching materials based on deep learning. Optik, 271, 170234.

[16] Yao, M., & Liu, J. (2024). The Analysis of Chinese and Japanese Traditional Opera Tunes With Artificial Intelligence Technology Based on Deep Learning. IEEE Access, 12, 21084-21091.

[17] Chen, X. (2023). The Influence of Traditional Opera Culture on the Modern Vocal Teaching. Applied Mathematics and Nonlinear Sciences, 8(2), 3456-3470.

[18] Xu, Y., Wang, W., Cui, H., Xu, M., & Li, M. (2022). Paralinguistic singing attribute recognition using supervised machine learning for describing the classical tenor solo singing voice in vocal pedagogy. EURASIP Journal on Audio, Speech, and Music Processing, 2022(1), 8.

[19] Krause, M., Müller, M., & Weiß, C. (2021). Singing Voice Detection in Opera Recordings: A Case Study on Robustness and Generalization. Electronics, 10(10), 1214.

[20] Mimilakis, S.I., Weiss, C., Arifi-Müller, V., Abeßer, J., & Müller, M. (2020). Cross-version Singing Voice Detection in Opera Recordings: Challenges for Supervised Learning. In Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019, vol 1168, pp. 429-445.

[21] Zhang, W., et al. (2022). Construction of AI Environmental Music Education Application Model Based on Deep Learning. Computational Intelligence and Neuroscience, 2022, 9423953.

[22] Agarwal, J., et al. (2022). pyAudioProcessing: Audio Processing, Feature Extraction, and Machine Learning Modeling. SciPy Conference Proceedings, 2022, 45-52.

[23] Dufourq, E., et al. (2024). Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks. Ecological Informatics, 75, 102457.

[24] Liu, H., et al. (2025). The evaluation model of engineering practice teaching with complex network analytic hierarchy process based on deep learning. Scientific Reports, 15, 14733.

[25] Kumar, A., et al. (2024). A Time-Distributed CNN-LSTM with Attention Model for Speech Based Emotion Recognition. Proceedings of the 2024 7th International Conference on Digital Medicine and Image Processing, 45-52.

[26] Li, J., et al. (2024). Multi-layer CNN-LSTM network with self-attention mechanism for robust estimation of nonlinear uncertain systems. Frontiers in Neuroscience, 18, 1379495.

[27] Zhang, L., et al. (2022). Music Emotion Classification Method Based on Deep Learning and Explicit Sparse Attention Network. Computational Intelligence and Neuroscience, 2022, 9239758.

[28] Prechelt, L. (1998). Automatic early stopping using cross validation: quantifying the criteria. Neural Networks, 11(4), 761-767.

[29] Zhang, X., et al. (2025). Optimization and Application of Teacher Performance Evaluation Model under the Background of Big Data. Proceedings of the 2025 2nd International Conference on Informatics Education, 234-242.

[30] Chen, H., et al. (2025). Deep learning-based strategies for evaluating and enhancing university teaching quality. Measurement: Sensors, 33, 100025.

[31] van der Kleij, F.M., et al. (2022). Personalized feedback in digital learning environments: Classification framework and literature review. Computers and Education Open, 3, 100080.

[32] Su, Y.S., et al. (2021). Personalized Online Learning Resource Recommendation Based on Artificial Intelligence and Educational Psychology. Frontiers in Psychology, 12, 767837.

[33] Patil, S.A., Pradeepini, G., & Komati, T.R. (2023). Novel mathematical model for the classification of music and rhythmic genre using deep neural network. Journal of Big Data, 10, 108.

[34] Liu, C., & Guo, W. (2025). Effectiveness of AI-Driven Vocal Art Tools in Enhancing Student Performance and Creativity. European Journal of Education, 60(2), e70037.

[35] Chang, D. (2025). Vocal performance evaluation of the intelligent note recognition method based on deep learning. Scientific Reports, 15, 13927.

[36] Kim, H.G., et al. (2018). Comparative study of singing voice detection based on deep neural networks and ensemble learning. Human-centric Computing and Information Sciences, 8(1), 35.

[37] Yang, Z. (2025). University English teaching evaluation using artificial intelligence and data mining technology. Scientific Reports, 15, 16498.

[38] Li, C. (2025). The integration and innovative practice of intelligent AI and local opera in college teaching. Frontiers in Psychology, 15, 1521777.

[39] Crawford, M. (2025). The impact of generative AI on school music education: Challenges and recommendations. International Journal of Music Education, 43(1), 45-62. https://doi.org/10.1177/02557614241308522

[40] Almubarak, A., et al. (2025). An AI-powered framework for assessing teacher performance in classroom interactions: a deep learning approach. Frontiers in Artificial Intelligence, 8, 1553051.

[41] Li, Y., et al. (2025). A teaching quality evaluation framework for blended classroom modes with multi-domain heterogeneous data integration. Expert Systems with Applications, 255, 125064.

[42] Li, H., et al. (2025). Adaptive deep reinforcement learning for personalized learning pathways: A multimodal data-driven approach with real-time feedback optimization. Computers and Education Open, 6, 100103.