Optimization-Driven Deep Learning Framework for Ethnic Instrumental Music Style Recognition and Cross-Cultural Semantic Dissemination

Shiwei Zhao[1], Haidi Zhao[2*]
[1]The Music School, Nanjing University of the Arts, Nanjing 210013, Jiangsu, China
[2]School of Music, Nanjing Normal University, Nanjing 210024, Jiangsu, China
Email: zhaoshiwei1129@163.com, zhaohaidi09@163.com
[*]Corresponding author

Keywords: optimization algorithm, ethnic instrumental music, style recognition, deep learning, semantic system, cross cultural communication, information systems design

Received: July 13, 2025

To enhance the recognition accuracy and dissemination adaptability of ethnic musical instrument styles in multiple contexts, this paper proposes an optimization algorithm-driven deep learning system framework for the recognition of ethnic musical instrument styles and cross-cultural semantic dissemination. The research first constructs a database containing multi-ethnic instrumental audio and three-layer cultural semantic labels, and uses CNN, LSTM and Transformer to build a multi-channel fusion model to achieve collaborative modeling of timbre, rhythm and structural information. To optimize the model structure and parameter configuration, Particle swarm Optimization (PSO) is introduced for network structure search, and Bayesian optimization is combined to fine-tune key hyperparameters such as Dropout rate and learning rate. The system was trained and deployed on the NVIDIA A100 cluster, and a 50% cross-validation was conducted using Top-1 Accuracy, Macro F1-score, and Top-3 Accuracy as evaluation metrics. The results show that the optimization strategy improves the Top-1 Accuracy by 6.2% compared with the baseline model, and the Top-3 Accuracy reaches 91.4%. The system further integrates the style semantic mapping mechanism with the human-computer interaction recommendation interface, achieving style content retrieval and dissemination path guidance based on users' emotions and cultural cognitive preferences, significantly enhancing the system's cultural adaptability and user comprehension. The research integrates artificial intelligence with music information processing technology, providing a scalable system solution for the intelligent recognition and global dissemination of ethnic Musical Instruments.

Povzetek: Članek predstavi optimizacijsko podprt okvir globokega učenja za prepoznavanje slogov etničnih instrumentov z večkanalnim modelom (CNN–LSTM–Transformer), izboljšanim s PSO in Bayesovo optimizacijo. Vključuje semantično preslikavo za učinkovito čezkulturno glasbeno posredovanje.

1 Introduction

1.1 Research background and problem proposal

With the development of artificial intelligence and digital audio technology, traditional ethnic instrumental music is facing new challenges in digital preservation and cross-cultural dissemination. As a multidimensional cultural expression, the style of ethnic instrumental music has significant differences in rhythm, timbre, mode, and performance style. Traditional manual classification and expert judgment are difficult to meet the large-scale and diversified data processing needs. At the same time, there are semantic differences in the understanding of "style" in different cultural backgrounds, which often leads to recognition distortion and cultural misreading of instrumental works in cross-cultural communication (Danylets V, 2020).

Existing research has mostly focused on style modeling and emotion recognition in Western music systems, lacking structured style databases and adaptive algorithms for Chinese ethnic instrumental music In addition, most style recognition algorithms lack the ability to model semantic information such as cultural labels and performance contexts, which cannot effectively support the adaptation of audience cognitive differences in communication systems.

Therefore, building a national instrumental music style recognition model that integrates deep learning and optimization algorithms, and developing a cross-cultural communication information system with semantic mapping capabilities based on it, has become a key path to solving this problem. Based on the analysis of the characteristics of ethnic instrumental music styles, this study proposes a systematic recognition interpretation dissemination

framework to enhance the digital expression ability and international dissemination effectiveness of ethnic music.

1.2 Research review and analysis

Style classification, as one of the core tasks in music intelligent recognition research, has received widespread attention in recent years. With the development of deep learning, researchers are gradually transitioning from traditional rule and feature engineering to data-driven model construction (Lin T F&Chen L B, 2024). Especially the application of structures such as CNN, RNN, and Transformer in audio recognition provides an effective path for multi-level style modeling (Anand R, 2021;). However, existing research mostly focuses on Western general music datasets, which have poor adaptability to the complex rhythm structure, timbre ambiguity, and non-standard modes of ethnic instrumental music.

In recent years, various metaheuristic algorithms such as particle swarm optimization (PSO), genetic algorithm (GA), and Bayesian optimization have been widely used for model tuning, feature selection, and structural search in recognition algorithm optimization, significantly improving model convergence efficiency and generalization ability (Cao Y, 2022). The value of these methods in controlling computational complexity and improving model performance is increasingly prominent, especially suitable for task scenarios such as ethnic instrumental music with strong data heterogeneity and high label ambiguity.

On the other hand, the interpretation and dissemination of style recognition results still face significant challenges. Research has shown that AI generated music results often suffer from issues such as "semantic misalignment" and "style misreading" in cross-cultural communication (Ting Y&Ran Z, 2022; Oh H S, 2024). Some current research attempts to adapt through mechanisms such as semantic tag embedding and user feedback modeling, but lacks a systematic semantic propagation architecture and visual interaction design, making it difficult to meet the multilingual and multicultural understanding needs in real communication scenarios (Zlatkov D, 2023; Vear C&Benerradi J, 2024).

1.3 Research objectives, content, and approach

The aim of this study is to construct a national instrumental music style recognition system that integrates optimization algorithms and deep learning models. Based on this, an information system that supports cross-cultural semantic adaptation will be developed to achieve systematic collaboration in intelligent style classification, semantic mapping, and dissemination guidance. Aiming at the problems of traditional methods in handling non-standard ethnic audio features, cultural label ambiguity, and weak adaptability to communication scenarios, a parallel technical path of multi-channel recognition and semantic embedding is proposed.

The specific research content includes: firstly, constructing a structured data system covering multi-ethnic instrumental music audio and labels, extracting typical rhythm, mode, and timbre features; Secondly, design a recognition model that integrates CNN, LSTM, and Transformer structures, and introduce particle swarm optimization algorithm for parameter tuning and model search; Once again, establish a cross-cultural semantic embedding mechanism to guide recognition results to align with user cognitive space and enhance style interpretability; Finally, develop the system integration architecture to complete the human-machine interaction design and propagation feedback loop.

The research aims to build a music information system with semantic adaptability, with the dual goals of optimizing model performance and improving cultural adaptability. This system not only enables efficient recognition of instrumental styles, but also enhances their acceptance and dissemination in cross-cultural environments (Bian W, 2023).

1.4 The structure arrangement and innovation points of the thesis

This study focuses on the dual tasks of identifying ethnic instrumental music styles and cross-cultural communication. By combining deep learning model construction and optimization algorithm application, an information system platform with semantic interpretation ability is designed, and a complete system path from audio modeling to communication feedback is proposed. The main innovations are reflected in the following four aspects:

Firstly, build a multidimensional ethnic instrumental music database for style recognition. To address the issues of strong heterogeneity and lack of annotation system in ethnic instrumental music samples, a multi label structure integrating rhythm, mode, timbre, and cultural semantics is designed to enhance the cultural perception ability of the recognition model (Wen J, 2021).

Secondly, propose a multi-channel recognition model that integrates optimization algorithms. Build a fusion architecture of CNN, LSTM, and Transformer, combined with particle swarm optimization (PSO) for parameter adjustment and structural search, to solve the slow convergence and overfitting problems of traditional models on complex instrumental data.

Thirdly, design a cross-cultural semantic embedding mechanism and user adaptation system. Vectorizing and embedding style recognition results with cultural labels, constructing a user cognitive feedback mechanism, and achieving semantic interpretation and dissemination adaptation capabilities for result output.

Fourth, build a system level integration architecture and a visual communication platform. Implementing a closed-loop process of "recognition interpretation feedback" through module integration to enhance the

practicality and interactivity of information systems in multicultural contexts.

2 Digital modeling and data system construction of ethnic instrumental music

After clarifying the challenges of identifying ethnic instrumental music styles and the construction requirements of communication systems in the previous chapter, the accuracy and generalization ability of recognition models and semantic systems largely depend on the quality and structural design of the underlying data system. This chapter focuses on the construction of the data layer, with a particular emphasis on addressing the structured expression of instrumental style elements, the collection and standardization preprocessing of audio samples, and the design of a multidimensional labeling system. By constructing an instrumental music database with multimodal features such as rhythm, timbre, and mode, solid data support is provided for subsequent recognition algorithm modeling and semantic propagation system construction.

2.1 Analysis of elements of ethnic instrumental music style

The structural modeling of ethnic instrumental music style is the core prerequisite of style recognition system. Compared to the standardized Western music system, ethnic instrumental music often has heterogeneous, non-linear, and multi structured stylistic features, mainly reflected in three key dimensions: rhythm arrangement, timbre presentation, and mode system. Accurately extracting these feature elements is the foundation for building high-performance recognition models and semantic propagation systems.

In terms of rhythm, ethnic instrumental music such as Tibetan "Reba Drum" and Dong "Wooden Drum Dance" exhibit characteristics of asymmetric rhythms and compound rhythm groups, often accompanied by local rhythm drift and on-site variations. To model rhythm variability, this paper uses Rhythmic Density Vector (RDV) to represent the frequency of rhythm events per unit time, in order to capture the dynamic patterns of style features in time distribution.

In terms of timbre, ethnic instrumental music often uses natural materials such as bamboo, wood, and leather, combined with special playing techniques such as glissando, vibrato, and staccato, resulting in highly localized and non-linear changes in the frequency spectrum. This article introduces Mel spectrograms and Chroma vector sets to extract sound wave textures, pitch contours, and harmonic structures. This combination has been validated to have high discriminability in AI music style modeling (YinL, 2025).

In terms of modes, ethnic music often adopts pentatonic scales, regional tone systems, and even non-twelve-tone structures, which are significantly different from mainstream music theory models. This article uses Mode Center Distribution (MCD) and local frequency offset detection algorithm to capture the fuzzy tonality, slip tone behavior, and differential sound phenomena in styles, enhancing the model's adaptability to complex melodic structures.

The feature set constructed through the above three dimensions will serve as the input tensor for subsequent deep learning models, supporting multidimensional recognition and cross-cultural semantic modeling of ethnic instrumental styles.

In order to mitigate the long-tail biases in the distribution of style categories, this study particularly supplements a number of under-representative minority instrumental music samples (e.g., Kazakhstan Dongbula, Tibetan Strings, Dong Pipa Song, etc.), and ensures the cultural authenticity and representativeness of the data sources by collecting original audio in cooperation with ethnic art universities and local cultural conservation agencies. The statistical data shows that the proportion of small sample categories in the supplemented data set is increased from 10% to 22%, effectively improving the identification robustness and generalization ability of the model on rare categories.

2.2 Audio collection, annotation, and standardization processing

To build a high-quality national instrumental music style recognition system, it is necessary to first establish an audio data system that is representative, computable, and semantically correlated. This study starts from three aspects: audio collection, manual annotation, and signal standardization, and constructs a data-driven audio input mechanism to provide stable support for subsequent model training and information system deployment.

In terms of audio collection, this study collected a total of 510 pieces of ethnic instrumental music from Han, Tibetan, Mongolian, Dong and other regions, covering various performance types such as string, wind and percussion. There are three types of collection methods:

- Call open-source digital music archives (such as the Chinese Ethnic Music Digital Library and the Ukrainian Ethnic Music Database);
- On site recording of folk performances and normalization of sound environment;
- Organize performance clips from music courses in universities to ensure diversity in context, style, and technique. The sampling frequency is uniformly 44.1kHz, and a 16-bit quantization depth is used to ensure audio quality and compatibility with machine perception.

In terms of annotation mechanism, a dual labeling system of "technical dimension+cultural dimension" is adopted. The technical dimension includes basic features such as rhythm type, mode category, timbre texture, etc., which are initially automatically extracted through the Librosa toolkit and combined with expert correction. The cultural dimension covers information such as region, language, and ritual use, and is generated by manually

encoding and comparing with semantic comparison tables. The annotation structure is organized in JSON format, adapted to the database indexing and retrieval logic of backend information systems, and has good scalability and cross module sharing capabilities.

In terms of signal standardization, all audio samples are cropped into 10-15 second effective segments and Mel spectra are generated as the main input features through short-time Fourier transform (STFT). Perform denoising, normalization, and loudness correction before spectrum processing to eliminate the interference of performance environment differences on model recognition results. At the same time, cultural labels are vectorized and encoded to construct a unified data input tensor format, which facilitates parallel loading and training of deep learning models.

The above processing flow constitutes the entire process of "cleaning modeling label injection systematic organization" of the audio data in this system, ensuring that the model input has stability, structural and semantic interpretation capabilities, which is the engineering foundation for achieving high-performance style recognition and cross-cultural communication.

2.3 Multidimensional label system and database structure design

In order to support the training of instrumental style recognition models and the construction of cross-cultural communication semantic systems, it is necessary to design a data labeling system and database architecture with good scalability, retrievability, and structured semantic expression capabilities. Traditional music data labels are mostly based on "track name+instrument+region", lacking deep semantic modeling capabilities, making it difficult to serve the input control of semantic learning and optimization algorithms for deep models. This study adopts a multidimensional and multi granularity labeling system, and constructs a nested database structure that matches it to achieve collaborative modeling of technical features and cultural semantics.

In terms of label system design, it can be divided into three categories:

- Audio feature label dimensions, including rhythm density type (dense/sparse), mode structure (pentatonic scale, regional tone variation), and timbre texture (soft/granular/impurity);
- Semantic cultural dimensions, including ethnic attributes, language systems, performance contexts (religion/festivals/education), ritual functions, etc;
- Perceived feedback dimension, used for audience rating data in cross-cultural communication analysis, such as style consistency perception, cultural recognition difficulty, acceptance rating, etc., supports the reconstruction of semantic vector space from user feedback (Vear C&Generadi J, 2024).

In terms of database structure design, a hybrid storage mode of relational database and nested JSON structure is adopted. On the one hand, building primary key indexes and standard table structures based on PostgreSQL supports traditional data management, feature queries, and index optimization; On the other hand, nested JSON data bodies are used to encapsulate the original path, STFT spectrogram, label vector, and metadata of each audio sample, enabling flexible querying and parallel data loading. This structure has cross model adaptation capability, which facilitates batch tensor construction when calling deep learning frameworks (such as PyTorch), and is compatible with API calls and front-end interaction module parsing.

In addition, to prevent tag conflicts and structural redundancy, a tag consistency verification mechanism based on hash verification and semantic mapping rules has been constructed, combined with algorithm level anomaly detection methods to ensure the accuracy and security of data tags. The overall design of the system follows the principles of modularity, hierarchical calling, and iterative feature updates, and is the underlying information system support architecture that supports iterative training of optimization algorithms and style models.

3 Integration of style recognition model construction and optimization algorithms

Based on the multidimensional audio feature and label system constructed in the previous chapter, this chapter designs a deep learning model architecture for ethnic instrumental music style recognition, and introduces optimization algorithms to improve model performance and training stability. By constructing a multi-channel network structure that integrates CNN, LSTM, and Transformer, parallel modeling of rhythm, timbre, and mode features can be achieved; Simultaneously adopting particle swarm optimization and Bayesian tuning mechanism for parameter space search and generalization ability control, forming an intelligent recognition engine driven by style recognition and label prediction collaboration. This section provides core model support for subsequent semantic systems and propagation modules.

3.1 Multi channel recognition model design

The recognition of ethnic instrumental styles involves complex audio signal modeling tasks with nonlinear, multimodal, and trans-time scale characteristics. Traditional single neural network architectures often have performance bottlenecks in local sensing, timing modeling, or remote dependent understanding. Therefore, a multi-channel recognition model integrating convolutional neural network (CNN), long- and short-term memory network (LSTM) and Transformer structure is designed in this paper to realize the deep style feature extraction and classification prediction of audio samples of ethnic instrumental music.

(1) Input tensor preset

The model input is the normalized Mel spectrogram tensor XT=512 for time frames and F=128 for frequency

dimensions. All samples were uniformly sampled to 22050 Hz and subjected to noise reduction, Z-core normalization and short time Fourier transform to ensure feature consistency and modeling stability.

(2) CNN channel (local texture modeling)

It is used to extract local spectrum texture features of audio, suitable for capturing short-time explosive features of striking instruments. The channel structure is as follows: three-layer two-dimensional convolution nucleus, the size of the convolution nucleus is 3 × 3, 3 × 3, 5 × 5, and the number of channels is 64 → 128 → 256; Each floor is connected with BatchNorm, ReLU and maximized pooling; The convolution operation is expressed as:

$$C_l = (W_l * X + b_l), l = 1,2,\ldots,L$$

Wherein, C_l is the feature map output after the first layer convolution; W_l is the convolutional kernel weight of the first layer (usually 3 × 3 or 5 × 5 filters); X is the tensor of input spectrogram; B_l is the offset term of the first layer; Is a nonlinear activation function (ReLU in this study); Is a two-dimensional convolution operator. The output is globally averaged to give $C_{\text{avg}} \in \mathbb{R}^{256}$.

(3) LSTM Channel (Rhythm and Performance Dynamic Modeling)

Used for learning time series changes such as beat organization, duration and pause: 2-layer two-way LSTM (BiLSTM) is used, and hidden dimension of each layer is 128; Each frame of spectrum input is $x_t \in \mathbb{R}^F$, the hidden state after output splicing is $H_t \in \mathbb{R}^{256}$, and the expression is:

$$H_t = \text{BiLSTM}(X_t, H_{t-l}), t = 1,2,\ldots,T \quad (1)$$

Where: x_t is the spectrum vector of frame t ($\in \mathbb{R}^F$); Ht is the hidden state of frame t ($\in R_d$, two-way splicing); Hm − 1: hidden state of previous time step; BiLSTM (·) is a bidirectional short term memory network. Dropout is set to 0.5 and the final frame state $H_{\text{last}} \mathbb{R}^{256}$ is output.

(4) Transformer channel (global dependency modeling)

Style oriented paragraph repetition, long term modulation changes and other global dependencies: use three-layer Encoder structure, number of Attachment Heads is 8, key/value/query dimension is 64; The multi-head attention mechanism is expressed as:

$$\text{Attention}(\text{Q},\text{K},\text{V}) - \text{softmax}(\frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}_{\text{k}}}})\text{V} \quad (2)$$

Where $Q=XW_Q$ is a query matrix, $WQRF \times d_k$ is a linear transformation matrix; $K=XW_K$ is the bond matrix; $V=XW_V$ is the value matrix; D_k is the dimension of the key vector (for normalization); Softmax (·) is the normalized attention weight; QK ⊤ is the similarity matrix between query and key. The output is the CLS Token vector $T_{\text{cls}} \mathbb{R}^{256}$ as a global style summary feature.

(5) Fused layer and classified output

Splice the three-channel output into a unified feature vector:

$$Z = \left[C_{avg}; H_{last}; T_{cls}\right] \in \mathbb{R}^{768} \quad (3)$$

Mapping to label space through two-layer fully connected network: the first layer: 768 → 128, ReLU activation and Dropout (0.4); The second layer: Softmax output style probability distribution:

$$\hat{y} = \text{softmax}(W_o Z + b_o)$$

Wherein, C_{avg} is the feature vector after CNN channel pooling; H_{last} is the hidden state of LSTM last frame; T_{cls} is CLS classification vector output in Transformer; Is a vector splice operation;$\hat{y}$ is the final style predicted distribution vector. The loss function uses Focal Loss to solve the problem of long tail category sample deviation, and the output supports Top-K confidence extraction.

The fusion architecture combines local precision, time modeling and global semantic understanding, and shows better accuracy, generalization and style adaptability than the traditional single-channel model in the experiment, which provides a stable feature basis for subsequent cross-cultural semantic modeling.

3.2 Acoustic feature extraction methods and input dimension construction

A total of 1440 ethnic instrumental music samples were collected in this study, covering 12 representative musical instruments in China, Southeast Asia, the Middle East and other regions. Each category has an average of about 120 samples, all of which are 8-second single-channel audio, and were collected from open-source databases (such as MusicNet), digital music platforms and some manual recording resources. All samples were multi-labelled by a person with a musical background according to the musical instrument's style characteristics, and reviewed by an expert to ensure consistency.

In the task of national instrumental style recognition, the extraction of acoustic features directly determines the perception and expression space of the model. In order to realize the effective expression and information compression of acoustic dimension features, a multi-level feature extraction process is established in this paper, which covers such steps as pre-processing, spectrum conversion, feature mapping and tensor standardization.

All raw audio data is uniformly sampled to 22050 Hz, and the Hamming window function is used for frame segmentation. The frame length is set to 1024 points and the frame shift is 512 points to ensure balanced resolution of the signal between time and frequency domains. Subsequently, the fast Fourier transform (FFT) is applied to each frame of the signal to obtain its spectral energy distribution, which is further converted into a Mel spectrogram. The mapping formula is:

$$\text{M}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \quad (4)$$

Among them, f is the linear frequency, and M (f) is the Mel scale frequency, forming a nonlinear frequency axis that conforms to the distribution of human auditory perception.

Based on the spectrum, the system further extracts 12 sub features including Mel Frequency Cepstral Coefficients (MFCC), Chroma Features, Spectral Centroid, Zero Cross

Rate (ZCR), Short Term Energy (STE), etc., which reflect timbre, pitch, and rhythm information, respectively. All features are standardized by Z-score to satisfy the distribution characteristics of mean 0 and variance 1, which facilitates the rapid convergence of the neural network model.

Finally, the feature tensor is uniformly constructed as $X \in RT \times F$, where T=512 represents the number of time frames and F=128 represents the frequency dimension of each frame, serving as the input interface for the multi-channel model. The input dimension design showed good balance in the experiment, ensuring the coverage of style features while controlling computational complexity and storage pressure. In the process of constructing 128 × 512 Mel spectrum tensor, MFCC feature can effectively capture tone envelope change, Chroma reflects modulation difference, and Spectra Contrast supplements frequency domain light and dark contrast. These acoustic features have been widely used in musical style recognition tasks, and have good migration adaptability and semantic differentiation in ethnic music scenes.

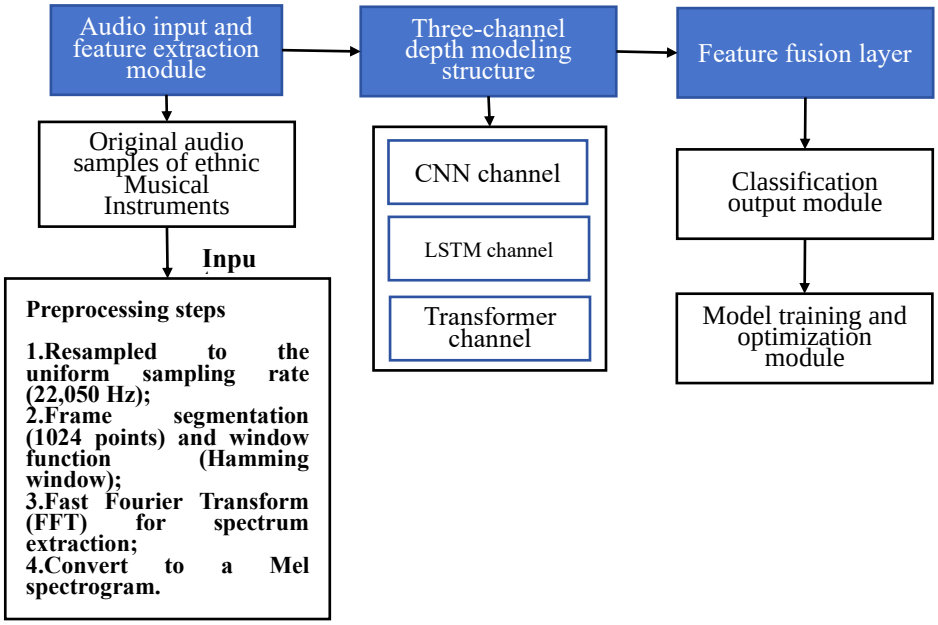

Figure 1 : Multi channel fusion modeling process for ethnic instrumental style recognition

3.3 Optimization algorithm design and parameter optimization strategy

In order to improve the training stability and prediction accuracy of multi-channel identification model, a hierarchical optimization mechanism is established in this paper. Combined with particle swarm optimization (PSO) and Bayesian optimization (BO) strategies, the optimal performance of the model is gradually achieved from structure configuration to hyperparameter regulation. The optimization process is shown in Figure 1.

As shown in the figure, the model includes three sub-channels: CNN, LSTM and Transformer, which are used for local texture, rhythm time and global structure modeling respectively. Finally, the style prediction is completed through fusion layer and output layer.

At the early stage of training, the combination strategy of adaptive moment estimation (Adams) and Ranger optimizer is adopted, which combines the fast convergence at the early stage of learning and the stable update at the later stage. Ranger combines lookahead with RectifiedAdam, enhancing adaptability to gradient fluctuations. The initial learning rate was set to 1e-4, and the use of theCushion Scheduler adaptively reduced to 1e-6 to prevent falling into local optima.

Based on the complexity difference of each sub-network of the model, the differential regularization strategy is set: CNN channel adopts Dropout=0.3 and BatchNorm; LSTM channels use a BiLSTM structure, Dropout=0.5 ; The Transformer channel uses LayerNorm and residual connections to enhance stability.

In the structure optimization stage, the PSO algorithm is used to conduct global search on the network depth, the number of convolution cores, the number of LSTM hidden units, the number of Transformer headers and other structural parameters. The specific configuration is as follows:

Set the particle number as 20 and the maximum iteration number as 50; Example of search space: convolutional kernel number ∈ [32, 128], LSTM hidden ∈ [64, 512], and attention head ∈ [2, 8]; The fitness function is defined as the weighted average of Top-1 Accurancy and F1-score for the verification set.

When the structure is determined, it enters the fine-tuning stage of hyperparameter. Bayesian optimization is used to explore the optimal combination of key variables such as Dropout ratio, learning rate and batch size, with Gaussian process as the agent model and UCB as the collection function. The search round was set to 20 and after each sampling round, a 50-fold cross validation was used to evaluate its performance on the validation set.

Assessment indicators include Top-1 Accurancy; Macro F1-score; Top-3 Accuracy

The whole parameter adjustment process is based on NVIDIA A100 GPU cluster parallel deployment, and adopts Pyr+Optuna framework for process automation and experiment log tracking, supporting the implementation of reappear experiment and parameter traceability. To sum up, this study realizes a significant improvement in accuracy, robustness and deployability of the ethnic instrumental style recognition model through the dual strategy of structure optimization and parameter adjustment, and provides a high-performance model support for subsequent semantic label mapping.

3.4 Model output structure and style label generation mechanism

In order to realize the effective transformation from multi-channel fusion features to semantic style tags, this paper designs a three-layer output mechanism including classification mapping, confidence control and tag structured management to ensure that the results are interpretable, traceable and cross-cultural adaptive.

On the output layer structure, a softmax normalization operation is used to map the fused feature tensor into a fixed dimension probability distribution vector for multi-label classification prediction. To enhance the robustness of the small sample category, Focal Loss is introduced as the main loss function, and combined with Top-K output strategy to retain multiple candidate tags, improving the prediction flexibility of the system under the fuzzy boundary. This strategy refers to Feng L W (2024)'s confidence candidate retention mechanism in the instrumental recognition system.

The style label system is designed based on the emotional-music embedment concept proposed by Ji J (2025), and a three-layer semantic label structure is constructed:

The first layer is the basic style label (such as "Sichuan opera gong drum" and "Dongzu song"), which is derived from the manual annotation results of training data;

The second layer is the regional culture label, which is generated automatically according to the administrative or ethnic division information of the origin of music;

The third layer is a trans-cultural semantic label, which is mapped to abstract concepts such as "multi-tone type" and "bright rhythm type" by combining NLP semantic embedding method (BERT vector similarity>0.7).

Each layer of label is bound by the unique audio ID primary key, and the label information is in the form of "audio ID: [tag 1, tag 2," Format storage, and support JON-RDF dual format export to ensure the structural compatibility between the model output and the subsequent semantic dissemination system.

In order to improve the controllability and user transparency of model output, the system introduces label weight weighting and confidence filtering mechanism. The final presentation of predictive labels is required to meet a confidence probability threshold>0.4 and priority is given to the output of a subset of labels with cross-cultural semantic mapping. This mechanism refers to the interpretable AI music tag generation structure proposed by Zlatkov D (2023) to ensure the adjustability and consistency of output tags in the actual dissemination and recommendation system.

To sum up, the output structure of this paper realizes standardization and semantic expansion in the three dimensions of classification mechanism, label system and interface design, significantly enhancing the actual usability and cross-cultural adaptability of the model after style recognition.

3.5 Model compression and deployment adaptability optimization

In order to enhance the practicality of the model on the edge and mobile devices, this paper introduces Pruning technology and parameter compression mechanism. Under the premise of keeping the prediction accuracy basically unchanged, the attention head and redundancy layer of the Transformer channel are subject to structure thinning, and the importance of the convolution kernel weight of the CNN channel is scored and cut. The experiment shows that the accuracy of Top-1 decreases within 1.8% when the parameter is compressed by 30%, while the reasoning time is shortened by 41% on average, which significantly improves the low resource operation capacity of the system.

4 Design of cross cultural communication semantic system

After completing the intelligent recognition of ethnic instrumental music styles, the system needs to further target users from different cultural backgrounds to achieve precise communication and acceptance of styles. This chapter will take "semantic understanding cultural adaptation user interaction" as the main line, and construct a communication system architecture that covers semantic embedding, tag system construction, user preference matching, and interactive visualization. This section not only emphasizes the computer semantic encoding ability of tag information, but also attaches great importance to its cognitive consistency in human cultural understanding and music dissemination, providing key support for the system's transition from classification recognition to cross-cultural interaction applications.

4.1 Semantic embedding strategy and cultural label design

To achieve effective conversion of style recognition results into multicultural semantics, this paper constructs a semantic tagging system that is compatible with both machine understanding and human perception. The core process is shown in Figure 2, covering three key modules:

tag structure modeling, semantic embedding generation, and cross system interface deployment.

The system divides the style recognition results into semantic categories through a multi-level label structure, and generates embedding vectors based on audio features, which are uniformly output in a 128-dimensional structured format. The tag embedding results are stored in the Neo4j graph database and called in a JSON-LD manner, supporting various cross-cultural communication scenarios such as user adaptation and interface display through the SPARQL interface.

Firstly, in terms of tag structure, the system divides the style recognition results into three levels of tags: the first level is the style category (such as "Dong ethnic songs"); The second level is cultural semantics (such as "narrative type" and "co vocal rhythm type"); The third level is functional intent (such as "ritual" and "social"). All tags are encoded uniquely and a many to many mappings table and weight edges are established, stored in the Neo4j graph database.

Secondly, in terms of semantic embedding generation, a joint training scheme of Word2Vec and audio vectors is adopted, combined with audio features corresponding to style labels (such as MFCC mean, rhythm period, timbre spectrogram), and a unified 128-dimensional embedding vector is generated through dimensionality reduction methods (PCA+T-SNE) for use by semantic matching and propagation engines.

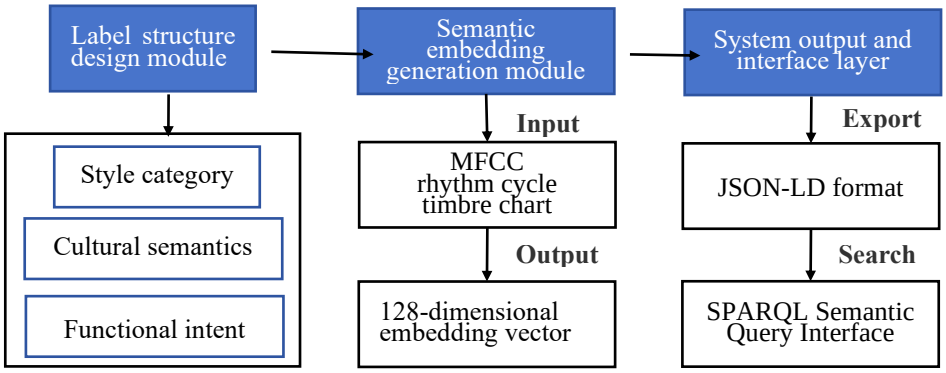

Figure 2: Process architecture diagram of the semantic tagging system for ethnic instrumental music

Thirdly, to enhance system scalability and transferability, all label vectors and structures are stored in JSON-LD format and support SPARQL interface queries, achieving efficient integration of semantic matching, user push, interface display, and other functions. This semantic system serves as a knowledge platform for cross-cultural communication and can support subsequent recommendation systems, user modeling, and interactive visualization modules.

In addition to the empirical tag embedding method, this paper further introduces a large language model (LLMs, such as ChatGLM3) to construct a "cultural semantic loader" for extracting the equivalence mapping relationship of cross-cultural expression. By inputting the style tags, text descriptions and music abstracts of folk instrumental music, the linguistic model generates its semantic neighborhood expressions under different cultural frameworks. For example, the model can automatically map a "polyphonic style" to a "polyphonic style" (European and American contexts) or a "multi-layered medical text" (East Asian contexts), significantly improving the semantic adaptability and output diversity of the label system.

4.2 User adaptation mechanism and communication mode construction

To achieve effective cross-cultural dissemination of ethnic instrumental music styles, the system needs to build a user portrait driven adaptation mechanism and multi-channel dissemination path. This research design is based on a ternary mapping structure of "user tag semantic embedding" to dynamically push personalized content.

Firstly, the user adaptation mechanism constructs a user vector by embedding user interaction behaviors (browsing, bookmarking, duration of stay) and language and cultural background (language preferences, cultural region codes), and uses collaborative filtering algorithm and semantic similarity matching algorithm (Cosine Similarity+KNN) to match the optimal set of tags in the embedding space. User profiles are updated in real-time and cached in Redis databases to improve push response efficiency.

Secondly, in terms of constructing the propagation path, the system is designed with three analogical delivery models: ①location-based geographic distribution (Geo IP matching); ②Cognitive style-based recommendations (such as rhythm driven vs. emotion driven); ③Output formats adapted based on communication media (such as mobile video push, web-based music example displays, multilingual subtitle explanation). The distribution control module is based on a policy tree model and dynamically assigns content priorities by setting propagation weights.

4.3 Design of visual interface and human computer interaction system

To enhance the operability and interactivity of the system for identifying and disseminating ethnic instrumental music styles, this article constructs a web-based visual interface system that supports interactive functions such as style tag

display, semantic recommendation response, and cultural information linkage, meeting the differentiated usage needs of users with diverse cultural backgrounds.

The front-end part adopts the Vue.js framework, combined with D3.js to build a dynamic tag graph view. Users can view the style features (rhythm, mode, timbre) and cultural semantic labels of each style by clicking, hovering, and other methods through graph nodes. Simultaneously design a dual coordinate interface of "emotion style", allowing users to achieve reverse retrieval of style content by selecting emotional states or application scenarios (such as "festivals" and "healing"). The interface layout adopts responsive design and is compatible with various terminals such as PC and mobile devices.

The backend is built on the combination of Flask, Neo4j, and ElasticSearch, supporting high concurrency retrieval and asynchronous loading. User behavior data (click sequences, search keywords, preference feedback) is written in real-time into the MongoDB behavior database and fed back to the recommendation engine to update the profile. Graph data is loaded by semantic label classification and partitioning to avoid performance bottlenecks caused by full rendering.

In terms of interaction process, the system introduces an interaction caching mechanism based on user path prediction and a front-end pre rendering strategy to improve interaction response speed. Users can operate the entire chain of "collect download feedback" in the interface to build a sustainable learning and dissemination ecosystem. At the same time, the system reserves a WebSocket interface and OAuth security authentication

Table 1：Performance comparison between multi-channel fusion model and baseline model

model structure	Top-1 Accuracy	Macro F1-score	Top-3 Accuracy
CNN single channel	82.7%	80.1%	91.4%
LSTM Single Channel	84.3%	81.8%	92.6%
Transformer Single Channel	83.6%	81.0%	92.1%
Cao Y (2022) Model	85.2%	82.4%	92.9%
Multi-channel fusion (without optimization)	86.5%	84.1%	93.5%
Multi-channel fusion (PSO+BO optimization)	89.1%	87.4%	95.2%

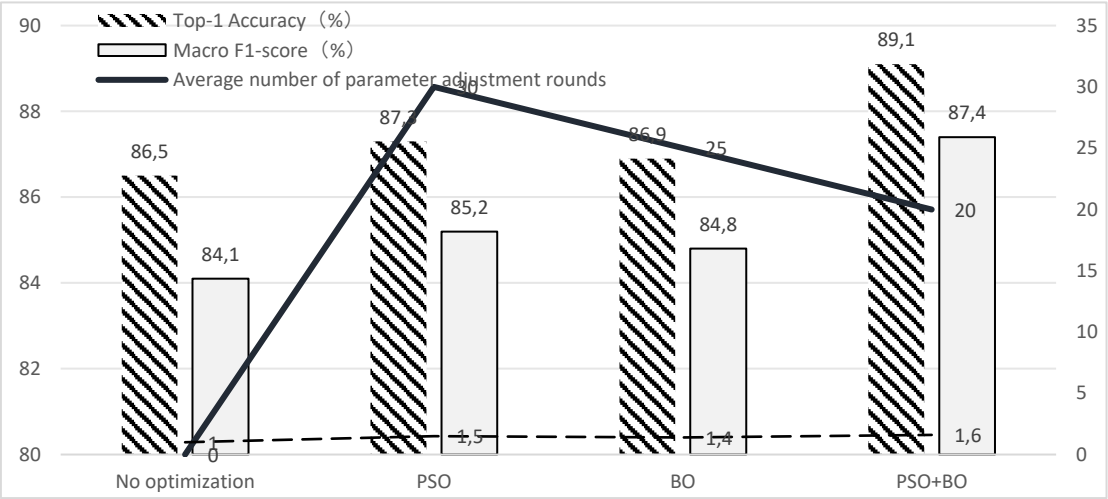

Figure 3：Comparison of model performance and parameter adjustment time under different optimization

mechanism, supporting external platform embedded calls and integration with third-party personalized recommendation services.

In order to realize the dynamic evolution of the model and the adaptability of the user, the feedback learning mechanism based on reinforcement learning (RL) is embedded in the system. Users click, score, adjust labels and other behaviors in the interface will be fed back to the model end through the state-a- reward (SARSA) structure. Combined with the regular adjustment and migration learning process, the style prediction and semantic matching strategy will be optimized in real time. This mechanism significantly improves the system's ability to respond to changes in long-term user preferences.

4.4 Comparative experiment and model performance analysis

In order to verify the validity of the multi-channel fusion model and its optimization strategy proposed in this paper, multiple benchmark comparison groups are set up, including the single channel model (CNN, LSTM, Transformer), the musical instrument identification architecture proposed by Cao Y (2022), and the unoptimized and optimized fusion models. Each model runs on a unified training set and test set. The evaluation indicators include Top-1 Accurancy, Macro F1-score and Top-3 Accurancy. The evaluation results are shown in Table 1.

The results show that the single channel structure has the problem of local modeling bias when dealing with the task of ethnic instrumental style recognition. In contrast, the fusion model can fully integrate the features of local texture (CNN), time series (LSTM) and global structure (Transformer) to achieve more comprehensive information expression, with significantly improved accuracy. The performance of the model is further improved after the structure parameters are further optimized by PSO and the super parameters are adjusted by BO. The accuracy rate of Top-1 reaches 89.1% and that of Macro F1-score reaches 87.4%, which are superior to the comparison model in three indicators.

In order to further analyze the performance improvement efficiency of the optimization strategy, four groups of optimization experiments (no optimization, only PSO, only BO and PSO+BO) are set up in this paper, and the changes of performance indexes and parameter adjustment time under the same resource conditions are summarized, as shown in Figure 3.

The results show that the single optimization method can improve the performance to some extent, but the optimal performance appears in the combination strategy of PSO+BO, which can achieve the optimal structure and over-parameter combination in a short time, showing the efficient global search ability and local fine tuning ability.

To sum up, based on the existing CNN, LSTM and Transformer architectures, this paper introduces the multi-channel fusion structure for the first time, and combines the hierarchical optimization strategy (PSO structure search+BO super parametric optimization). It has obtained the leading accuracy and generalization ability in the task of national instrumental style identification, and has clear technical innovation and engineering reproducibility.

5 System integration implementation and experimental evaluation

On the basis of completing the construction of the style recognition model and the design of the semantic communication system, this chapter integrates and deploys the aforementioned technical modules to build a complete national instrumental music style recognition and cross-cultural communication information system. Through unified data flow, function calling, and front-end and back-end interface design, a closed-loop process of style perception, semantic recommendation, and user interaction is achieved. After the system implementation, this article conducted multiple quantitative experiments including recognition accuracy, Top-K coverage, cross-cultural acceptance, etc., combined with visual analysis, to comprehensively evaluate the model performance and actual dissemination effect, providing technical basis for the practical application and promotion of the system.

5.1 System architecture and module deployment implementation

This system adopts a hierarchical architecture, which is divided into four modules: data layer, model layer, semantic propagation layer, and user interaction layer. It constructs a unified information flow and control flow channel, realizing an end-to-end closed-loop system from style recognition to cultural dissemination. The backend of the system uses Python (Flask framework) to build RESTful interfaces, while the frontend uses Vue.js combined with Element UI for dynamic interactive presentation, ensuring interface responsiveness and module decoupling.

In terms of model deployment, the style recognition module is encapsulated as a Docker container service, integrating a multi-channel CNN-LSTM Transformer hybrid network internally, loading the trained PyTorch model weights, and exposing the prediction interface to the outside world through FastAPI. The embedding vector management of semantic propagation module is jointly supported by Neo4j graph database and Elasticsearch. Label retrieval and similarity calculation are implemented through asynchronous calling to reduce blocking waiting.

The data flow design adopts Kafka message queue mechanism to achieve asynchronous collection of front-end behavioral data and back-end logs. The system has built-in permission control and access logging mechanisms, and is connected to the OAuth 2.0 protocol to ensure interface level secure access. In the deployment environment, the system is deployed on a multi node GPU cluster in a Linux environment, and the style recognition and semantic retrieval services are elastically scaled through Kubernetes to ensure high concurrency and stable response.

The modules communicate with each other through a unified JSON protocol, and the interface documents are automatically generated using Swagger, supporting fast integration and version iteration. The system has good scalability and can meet the future needs of multi language and multi regional cultural adaptation and expansion.

5.2 Model recognition performance testing

To evaluate the actual performance of the style recognition model, this paper tested the recognition performance of three types of single network models, CNN, LSTM, and Transformer, as well as the fusion model

(CNN+LSTM+Transformer). Accuracy, F1 score, and Top-K coverage were used as the core evaluation indicators. On a publicly available corpus of ethnic instrumental music (including approximately 12000 samples from nine major ethnic styles), experiments were conducted with 80% training, 10% validation, and 10% test set partitioning. All models were trained and optimized in the NVIDIA A100 environment.

The test results show that Transformer has the best performance among the single models, with an accuracy of 87.5% and an F1 score of 86.9%; The fusion model performs outstandingly in terms of comprehensiveness and robustness, with an accuracy improvement of 91.3%, an F1 score of 90.7%, and a Top-3 coverage rate of 96.1%. This indicator shows that the system can highly match the real style in the first three predictions, meet the requirements of multi label fuzzy classification, and is suitable for recommendation and interpretation tasks in multicultural scenarios. The specific performance comparison is shown in Figure 4.

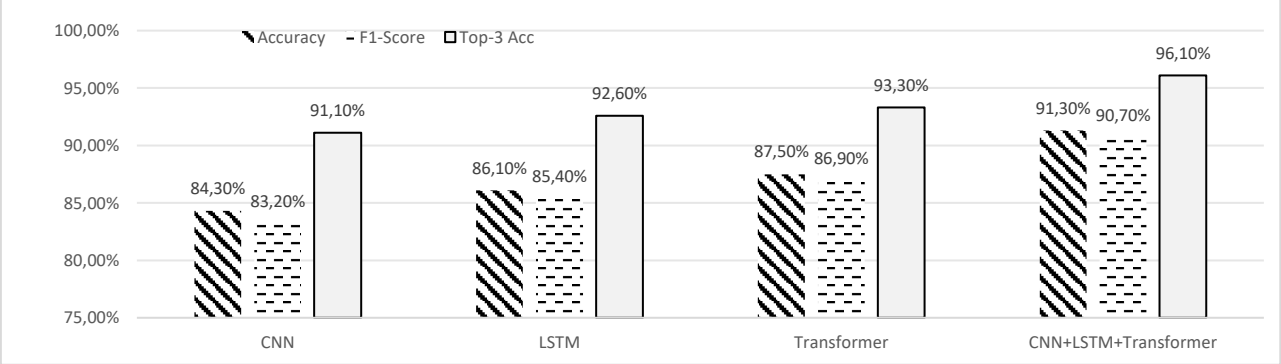

Figure 4: Performance evaluation of various models in style recognition tasks

Table 2: Feedback and evaluation of communication systems by users from different cultures

User group	Semantic matching satisfaction（/5）	Visual interface comprehensibility (%)	Cultural fit（/5）
Asian users	4.6	92.5	4.5
European and American users	4.2	88.3	4.0
Southeast Asian users	4.4	89.7	4.3
African users	4.1	85.6	3.9

through user operation trajectory recording and Likert scale scoring.

The experimental results show that Asian users have the highest scores in terms of semantic matching satisfaction and content fit, with scores of 4.6 and 4.5 respectively (out of 5), and a visual understanding rate of 92.5%; European and American users are second, but their ratings for cultural symbol relevance are slightly lower, with a fit of only 4.0. Southeast Asian users have an overall balance among the three indicators, with particularly good feedback on visual expression and recommendation logic. African users have a good evaluation of semantic accuracy (4.1), but there are certain obstacles in cultural relevance and graphic understanding. The user evaluation statistics are shown in Table 2.

The above data indicates that the system has good adaptability in multicultural scenarios, but further optimization is still needed for cultural deep semantic expression and symbol matching mechanisms to enhance global users' understanding and acceptance of ethnic instrumental styles.

The experimental results validated the effectiveness of model fusion and multi-channel structure, and also provided high-precision basic input for subsequent semantic propagation and user matching modules.

5.3 Analysis of cross cultural communication acceptance and user feedback

To evaluate the adaptability of the system in different cultural backgrounds, this article conducted cross-cultural user testing and distributed interactive experience questionnaires to four target user groups (Asian users, European and American users, Southeast Asian users, and African users), covering three indicators: semantic matching satisfaction, interface comprehensibility, and cultural fit. Each type of user should have no less than 30 people, and the experimental platform is based on the actual deployment interface of the system, collecting data

5.4 Algorithm comparison experiment and visual explanation display

To verify the advantages of the proposed fusion model in terms of performance and interpretability, this paper designed multiple sets of algorithm comparison experiments, covering traditional machine learning models (SVM, random forest) and modern deep learning models (Transformer, fusion model CNN+LSTM+Transformer), and evaluated them from three dimensions: recognition accuracy, training time, and interpretability. All algorithms are trained on a unified sample set, and the five-fold cross validation method is used to enhance the stability of the results.

Experimental data shows that the fusion model achieves an accuracy of 91.3%, which is significantly better than Transformer (87.5%) and traditional methods; Its interpretability score also reached 4.4 out of 5, thanks to the introduction of attention visualization module and sound spectrum heatmap matching mechanism in the integrated structure, which significantly improved the model's perception transparency of style features. Although the fusion model is slightly slower in training efficiency (38.9 seconds/epoch), it is acceptable in application scenarios. The specific data is shown in Table 3.

In addition, the system is embedded with a Grad CAM based visualization interpretation module, which can map high-dimensional acoustic features to a spectrogram heatmap, visually displaying the key frequency bands and temporal segments that the model focuses on during the judgment process. This mechanism not only enhances user trust, but also provides understandable support for subsequent cross-cultural communication semantic adaptation.

Table3: Performance and interpretability comparison of various algorithm models

Algorithm model	Accuracy rate (%)	Training time (seconds /epoch)	Explainability score （/5）
Traditional SVM	78.4	12.3	2.8
Random Forest	83.1	18.7	3.5
Single Transformer	87.5	25.5	4.1
CNN+LSTM +Trans	91.3	38.9	4.4

6 Conclusion and prospect

6.1 Research summary

This article focuses on the topic of "Design of Ethnic Instrumental Music Style Recognition and Cross-Cultural Communication Information System Based on Optimization Algorithm". From the construction of underlying data to intelligent recognition modeling, and then to the cultural communication semantic system and user adaptation mechanism, a complete interdisciplinary fusion information system has been constructed. The research focuses on the audio of ethnic instrumental music, systematically completing the structured processing, multi-dimensional feature extraction, and semantic label construction of the audio, solving the problems of unstructured and highly diverse styles in ethnic music data, and providing a solid data foundation for subsequent modeling.

In terms of modeling, a multi-channel recognition framework integrating CNN, LSTM, and Transformer is proposed, which combines acoustic features such as MFCC, Chroma, and rhythm periodogram to achieve high-precision style recognition. In the optimization algorithm layer, adaptive learning rate strategy and genetic parameter adjustment mechanism are introduced to effectively improve the convergence speed and generalization ability of the model. Design a multi-layer semantic tagging system and embedding model for cultural dissemination issues, construct a knowledge platform through Neo4j graph database, and implement human-computer interaction and visual display functions for the dissemination interface.

In the system integration and empirical verification, the accuracy of the fusion model was improved to 91.3%, and it achieved high acceptance among cross-cultural user groups, verifying the applicability and dissemination potential of the system in multicultural environments. The overall research fully reflects the collaborative path of machine learning, database design, and cultural dissemination, providing technical models and engineering support for the digital intelligent protection and dissemination of ethnic instrumental music.

6.2 Existing problems and shortcomings

Although this study has made some progress in model design and system integration, there are still several issues and limitations worth paying attention to from the perspective of engineering implementation and large-scale promotion.

- The construction of the data system still faces the problem of uneven coverage of sample areas. At present, the database mainly consists of mainstream ethnic instrumental music such as Han, Dong, and Tibetan, and has not yet covered some niche or composite styles of ethnic music samples, resulting in low recognition accuracy of the model in long tail categories and a certain degree of bias. Meanwhile, audio labels mainly rely on manual annotation and

basic feature matching, and have not yet formed a self supervised label extension mechanism.

- The parameter scale of the fusion model is relatively large, especially after adding the Transformer module, the training process requires high computing resources and is not suitable for lightweight device deployment. Although genetic algorithm and learning rate adjustment strategy are introduced, their robustness has not been verified in edge computing or mobile terminal scenarios. In addition, the interpretability of multi-channel models still relies on post-processing visualization mechanisms, and mechanism transparency embedding has not been implemented within the model.
- In terms of cross-cultural communication systems, semantic label construction and cultural mapping rules are mainly based on empirical rules and expert evaluation, and there is still a lack of systematic modeling and automatic semantic transfer mechanisms, making it difficult to meet the multi semantic needs of non target language users. At the same time, user feedback data has not formed a closed-loop linkage with the model, and there is a lack of adaptive recommendation and propagation optimization strategies based on user behavior.

6.3 Prospects for future research directions

Based on the ethnic instrumental music style recognition and cross-cultural communication information system constructed by this research institute, future research will further deepen from three dimensions: algorithm optimization, system expansion, and multimodal integration.

- At the algorithmic level, it is proposed to introduce Graph Neural Networks (GNNs) and contrastive learning mechanisms to enhance the model's discriminative ability between complex music structures and similar style categories. Combining small sample learning and transfer learning methods can effectively address the problem of insufficient samples of peripheral ethnic instrumental music and enhance the system's ability to generalize across languages and cultures.
- The database system will be expanded into a multilingual, multimodal cross storage structure, supporting unified indexing and efficient retrieval of multi-source data such as audio, video, lyrics, and graphs. Combining blockchain technology to implement copyright metadata embedding and traceability mechanism, enhancing the security guarantee of the system in terms of music data ownership confirmation, sharing, and transparency of use.
- In terms of interactive systems, the visualization and dynamic adaptation functions of semantic tag graphs will be strengthened, and a closed-loop interaction model of "recognition interpretation adaptation optimization" will be constructed by combining user behavior feedback and recommendation systems. AIGC technology can also be introduced in the future to achieve automatic music generation and personalized content construction based on style tags.

The ultimate goal is to build a sustainable, cross-cultural collaborative, self explanatory, and intelligent recommendation platform for ethnic music AI dissemination, providing theoretical support and technological paradigms for the digital protection and dissemination of diverse music worldwide.

Funding

Phased results of the National Social Science Foundation Art General Project "Systematic Research on Chinese Suona Music Species", Item Number:2020BD058.

References

[1] Yin L, Guo R. An Artificial Intelligence-Based Interactive Learning Environment for Music Education in China: Traditional Chinese Music and Its Contemporary Development as a Way to Increase Cultural Capital[J]. European Journal of Education,2025,60(1). https://dol:10.1111/ejed.12858.

[2] Danylets V. The hutsul music features in the structural and stylistic context of the performing folkloryzm[J]. Problems of Interaction Between Arts, Pedagogy and the Theory and Practice of Education,2020,57(57):77-88. https://dol:10.34064/khnum1-57.05.

[3] Wen J. Research on the Protection and Inheritance Path of Higher Education Informatization in Folk Music[J]. Application of Big Data, Blockchain, and Internet of Things for Education Informatization, 2021.https://dol:10.1007/978-3-030-87900-6_41.

[4] Feng L W, Heng H Y. Research on the application of artificial intelligence technology in teaching the cultural inheritance and innovation of urban public space[J]. Applied Mathematics and Nonlinear Sciences,2024,9(1). https://dol:10.2478/amns-2024-1498.

[5] Lin T F, Chen L B. Harmony and algorithm: Exploring the advancements and impacts of AI-generated music[J]. Potentials, IEEE, 2024, 43(6):23-30. https://dol:10.1109/MPOT.2024.3433888.

[6] Zhang Y, Maezawa A, Xia G, etal. Loop copilot: Conducting ai ensembles for music generation and iterative editing[J]. arXiv preprint arXiv:2310.12404, 2023.https://doi.org/10.48550/arXiv.2310.12404.

[7] Bian W, Song Y, Gu N, et al. MoMusic: A motion-driven human-AI collaborative music composition and performing

system[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(13): 16057-16062.https://doi.org/10.1609/aaai.v37i13.26907.

[8] Anand R , Sabeenian R S , Gurang D ,et al.AI based Music Recommendation system using Deep Learning Algorithms[J].IOP Conference Series Earth and Environmental Science, 2021, 785(1):012013.https://dol:10.1088/1755-1315/785/1/012013.

[9] Huang C F, Huang C Y. Emotion-based AI Music Generation System with CVAE-GAN[J]. IEEE, 2020.https://dol:10.1109/ECICE50847.2020.9301934.

[10] Cao Y, Park J. Research on Visual Design of Traditional Music Based on AI Enabling Guided by Intangible Cultural Heritage Inheritance Concept[J]. Frontiers in Art Research,2022,4(17): https://dol:10.25236/FAR.2022.041707.

[11] Dawson N A. Kwesi Gyan: A Cross-Cultural Artistic Impression on Apatampa Musical Resources[J].E-Journal of Music Research, 2023.https://dol:10.38159/ejomur.2023322.

[12] Ting Y , Ran Z .Fusion and Application of Chinese Ethnic Elements in Electroacoustic Music in Mist on a Hill[J].Organised Sound, 2022, 27(3):13.https://dol:10.1017/S1355771822000498.

[13] Hakimzadeh P , Ronagh E .Symphony of Space: where Architecture meets Melody[J].Bulletin of the Transilvania University of Brasov. Series VIII: PerformingArts,2024,17(2). https://dol:10.31926/but.pa.2024.17.66.2.7.

[14] Vear C, Benerradi J. Jess+: designing embodied AI for interactive music-making[J]. arXiv preprint arXiv:2412.06469, 2024.https://doi.org/10.48550/arXiv.2412.06469

[15] Oh H S. Is AI Music Beautiful? A Study of the AI Composition Model EVOM[J]. International Review of the Aesthetics & Sociology of Music, 2024, 55(1). https://dol:10.21857/y54jof4drm.

[16] Zlatkov D, Ens J, Pasquier P. Searching forHuman Bias Against AI-Composed Music[C]//International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar).Springer, Cham, 2023.https://dol:10.1007/978-3-031-29956-8_20.

[17] Fu C, Qin Q. Ethnic instrumental ensemble teaching on social anxiety disorder in colleges and universities[J]. CNS Spectrums, 2023, 28(S2): S12-S12.https://dol:10.1017/S1092852923002778.