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

AI-assisted cross-cultural delivery of traditional musical art forms in contemporary times

Yujia Yang[1], *Dan Shen*[2,*] *and Xuandong Sun*[3]

[1] College of Music, Luoyang Normal University, Luoyang 471934, Henan, China
[2] School of Art, South China University of Technology,Guangzhou,510006,China
[3] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou,510006,China

SUMMARY: *As a cultural subject, traditional music and art are facing the technical problem of incompatible conversion forms in the inheritance and development of modern society. In this paper, we take "music score recognition - music visualization - cross-cultural transmission" as the research idea, and focus on constructing the application framework of AI technology. The detailed features of musical notes are extracted using Harr wavelet transform, and the note contour layer is processed by 1D convolution operation, and a multi-scale feature fusion CRNN framework is established for sheet music recognition. After completing the transcription of musical notes, the VGGish model is used to process the music data, and the t-SNE algorithm is used to visualize the music data, forming a cross-cultural output scheme for traditional music art. In the practical application of this scheme, the effect of music performance was recognized by 80.00% and above of the subjects, and the visualization effect was positively evaluated by 67.27% of the subjects, which is an effective development path for the traditional music art form to fit into the contemporary society under the AI technology.*

KEYWORDS: *music score recognition; VGGish model; t-SNE; music visualization; traditional music*

1 Introduction

Traditional music is a valuable treasure of human civilization, carrying unique cultural connotations and historical memories, and forming a huge and rich system of art forms with unique oriental aesthetic qualities [1, 2]. In the context of globalization, these traditional music and art forms are experiencing unprecedented dissemination and exchange, which not only show the cultural diversity, but also become an important bridge to promote international understanding [3, 4]. However, unlike calligraphy, painting and other art forms with physical carriers, traditional music culture is more difficult to preserve, and many musical genres and pieces suffer from problems such as inheritance faults, decreasing audience groups, and narrow dissemination channels, which pose formidable challenges for the inheritance and dissemination of traditional music culture [5-8]. Artificial Intelligence (AI) technology can effectively revitalize the music cultural heritage, preserve music in digital form, and form digital resources to provide support for subsequent innovative applications and expansion.

Cao and Park [9] (2022) created an AI-based music visualization system for intangible

*shendan2025@163.com
https://doi.org/10.65102/is2026524
Published: 30 April, 2026

cultural heritage to enhance the visual comprehensibility of traditional music through 3D animated interactions and music games, expanding musical artistic expression and promoting its inheritance. Cao and Park [10] (2022) created an AI-based music visualization system for intangible cultural heritage to enhance the visual comprehensibility of traditional music through 3D animated interactions and music games, expanding musical artistic expression and promoting its inheritance. Bosi et al [11] (2024) proposed an AI-driven audio cultural heritage record preservation method based on AI extracting key information features in audio and its associated video, automatically correcting errors and creating audio copies, repairing the audio, and facilitating audio preservation. Kuremoto [12] (2024) recognizes the elemental features of guqin music, such as rhythm, harmony, and tempo, with the help of machine learning, and is able to extract musical symbols in the score to optimize the musical expression. Yao and Liu [13] (2024) constructed a multi-layer deep learning model of multiple neural networks for extracting the features of Chinese and Japanese traditional opera tunes, which were used to distinguish the tunes and further analyze the Chinese and Japanese operas, and were able to fix the music sources for the cross-cultural dissemination of traditional music. Chinnasamy et al [14] (2025) developed a music lyrics generator and translator using natural language processing techniques for generating coherent and original lyrics content, which can be converted between lyrics in multiple languages to facilitate cross-cultural communication. Wei [15] (2025) combines convolutional neural networks, which are used to extract audio sample features, and fuzzy logic, which captures user preferences, to generate personalized song recommendations, and the framework helps in composing, mixing and generating music. Bi [16] (2025) states that AI and online collaborative platforms enable musical innovation and cross-cultural forms of expression, while the combination of virtual and augmented reality technology promotes a musically immersive experience for the user, enabling music to cross geographical and cultural boundaries. However, Eraslan [17] (2025) emphasizes the ability of AI to not only mimic music, but also reshape musical memories, due to the lack of human emotion and cultural depth in AI-assisted traditional music, and the shortcomings of AI in traditional music delivery.

In this paper, we first summarize the construction idea of the improved CRNN music score recognition model under multi-scale features, analyze the operation principle and content of its four modules of downsampling, note contour refinement, note classification prediction and sample learning in turn, and establish the music score recognition model based on multi-scale feature fusion. Then the VGGish model is used as the processing method of music data, on the basis of which the t-SNE music data visualization method is designed to explore the value embodiment dimension of music visualization dynamic design. Subsequently set up the effect comparison and evaluation experiments of music score transcription to verify the operation effect of the music score recognition model. Output the results of data processing based on VGGish model and music visualization based on t-SNE algorithm to demonstrate the effectiveness of both in music data processing and downscaling visualization. Based on the music score recognition model, music emotion classification and feature importance ranking are performed. Finally, the experimental data of the model and algorithm are analyzed from two levels, namely, music performance effect and visualization effect, to check the overall feasibility and explore the development path of digital music visualization accordingly.

2 Music Score Recognition Model Based on Multiscale Feature Fusion

2.1 Downsampling module

A sampling strategy under the Haar wavelet transform is proposed to address two problems in the music transcription task: (1) the difficulty of extracting the subtle differences between notes and symbols, and (2) the possible overlap between the notes and the background of the pentatonic score.The Haar wavelet transform is a wavelet transform with less loss of spatial resolution and better information retention.

The base functions ϕ and λ are defined in the one-dimensional Haar transform, ϕ is the smoothing component, which is used to capture the structure and overall contour of the note, and λ is the detail component. The basis functions of different scales are constructed recursively to correspond to different levels of detail features, and the smoothing component is shown in Eq. (1):

$$\phi_1(x) = \frac{1}{\sqrt{2}}\varphi_{1,0}(x) + \frac{1}{\sqrt{2}}\varphi_{1,1}(x) \tag{1}$$

The detail function λ is obtained by differencing neighboring basis functions, increasing the sensitivity to differences in the subtle structure of notes in the score, in the form of equation (2):

$$\lambda_1(x) = \frac{1}{\sqrt{2}}\varphi_{1,0}(x) - \frac{1}{\sqrt{2}}\varphi_{1,1}(x) \tag{2}$$

Recursion in the basis function generates basis functions at different scales and locations by scaling and translation operations, adapting to notes of different sizes. For any integers j and k, the basis functions are as in equation (3), using j to control the scaling of the function and k to adjust the translation:

$$\varphi_{j,k}(x) = \sqrt{2^j}\varphi(2^j x - k), k = 0,1,\cdots,2^j - 1 \tag{3}$$

The Haar wavelet transform $\varphi_{0,0}(x)$ is a Haar basis function defined as 1 in the interval $[0,1)$ to represent the recognition of the note head.The localization property of the Haar basis function makes it possible to recognize the note head efficiently in complex musical scenes, reducing the interference of other notes or quintuplets. The form is shown in equation (4):

$$\varphi_{0,0}(x) = \varphi_0(x) = \begin{cases} 0, & x < 0 \text{ Or } x \geq 1 \\ 1, & 0 \leq x \leq 1 \end{cases} \tag{4}$$

The smooth component of the Haar transform preserves the overall framework and approximate note positions in the score, while the detail component is suitable for detecting detailed variations in these notes.The structure of the Haar feature extraction module is shown in Fig. 1.

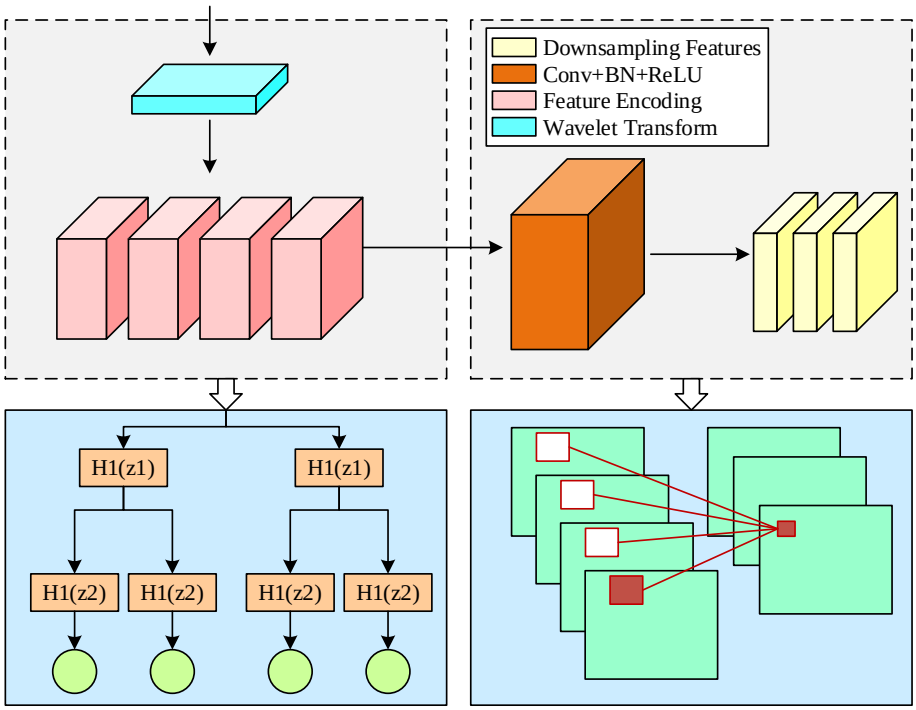


Figure 1: Haar feature extraction module architecture

2.2 Note Contour Refinement Layer

The structure of CCRL is shown in Fig. 2, which ensures the accuracy of music score detection for edge detection.CCRL first goes through a normalization layer, which normalizes the input feature maps to minimize the distributional differences between the features and improve the efficiency of the network training process.

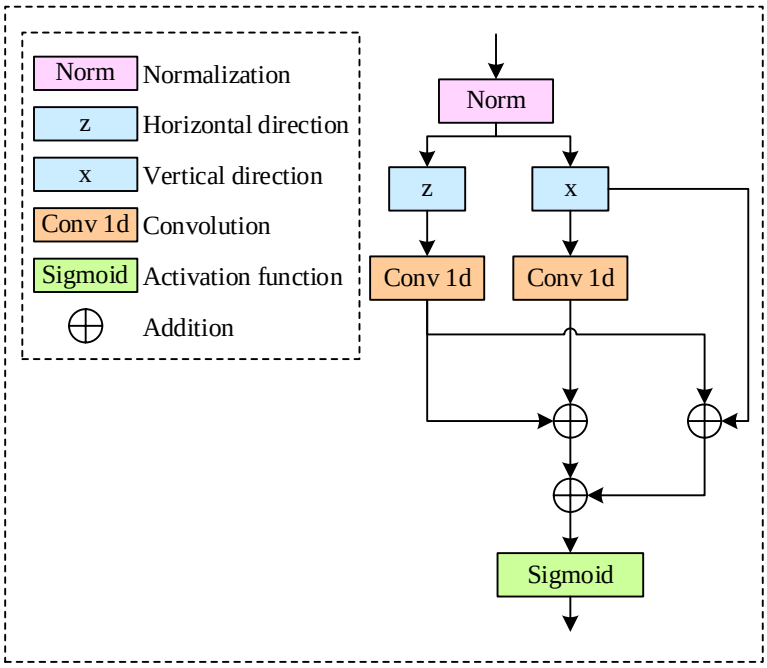


Figure 2: Detail layer of note contour

The normalized feature map is passed into two parallel 1D convolutional branches. The first branch of convolution is responsible for processing the horizontal direction of the feature map, and the second branch of convolution is responsible for processing the vertical direction of the feature map. The two branches use the 1D convolution to extract different local structures to capture the local detail structure of the note. The interacted feature maps are subjected to an addition operation, and the convolved and interacted feature maps are combined with the original input feature maps using residual concatenation.

2.3 SRU-based note classification prediction

To improve the training and inference speed, the gating mechanism in SRU is optimized and the gating parameter matrix is pre-computed. In the SRU model, x_t denotes the input data, and x_t is nonlinearly transformed by the Sigmoid activation function under the parameter vectors to get the new output state and internal state. h_t denotes the output state and c_t denotes the internal state. v_f and v_r are the parameter vectors for mapping c_{t-1}.

In the optimization recursion process, the calculation of the gating unit and the hidden state are adjusted. The hidden state no longer depends on the last time point h_{t-1}, but is based on the last internal state c_{t-1}. In this process, the forgetting gate is computed as in Eq. (5) and the reset gate is computed as in Eq. (6). b_f and b_r are the bias cells of f_t and r_t, respectively. The c_t combines the information from the past state and the current input, and its computation is reduced using Hadamard product as in equation (7):

$$f_t = \sigma\left(w_f x_t + v_f \otimes c_{t-1} + b_f\right) \tag{5}$$

$$r_t = \sigma\left(w_r x_t + v_f \otimes c_{t-1} + b_r\right) \tag{6}$$

$$c_t = f_t \otimes c_{t-1} + \left(1 - f_t\right) \otimes g_t \tag{7}$$

The h_t employs a jump connection to optimize the gradient transfer, which is computed in Eq. (8):

$$h_t = r_t \otimes c_t + \left(1 - r_t\right) \otimes x_t \tag{8}$$

2.4 Balanced Sample Learning Based on Focal Loss

Focal Loss is a loss function for the importance of extracting information from a small number of samples, which enables the neural network to assign balanced weights to the samples during row channel assignment. The overfitting and underfitting problems are mitigated after the model learns the inter-class balanced feature channels, reducing the misrecognition rate of a small number of sample class symbols.The Focal Loss loss function is defined in equation (9):

$$L_{Focal\_loss}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{9}$$

p_t is the prediction probability, α_t and γ are the adjustable factors, and $L_{Focal\_loss}(p_t)$

is the loss function when the prediction probability is p_t, which integrally measures the degree of matching between predicted sequences and real sequences and the difficulty of prediction.

3 Dimensionality Reduction and Visualization of Music Data

3.1 Data processing based on VGGish modeling

The modeling results for VGGish are shown in Figure 3.

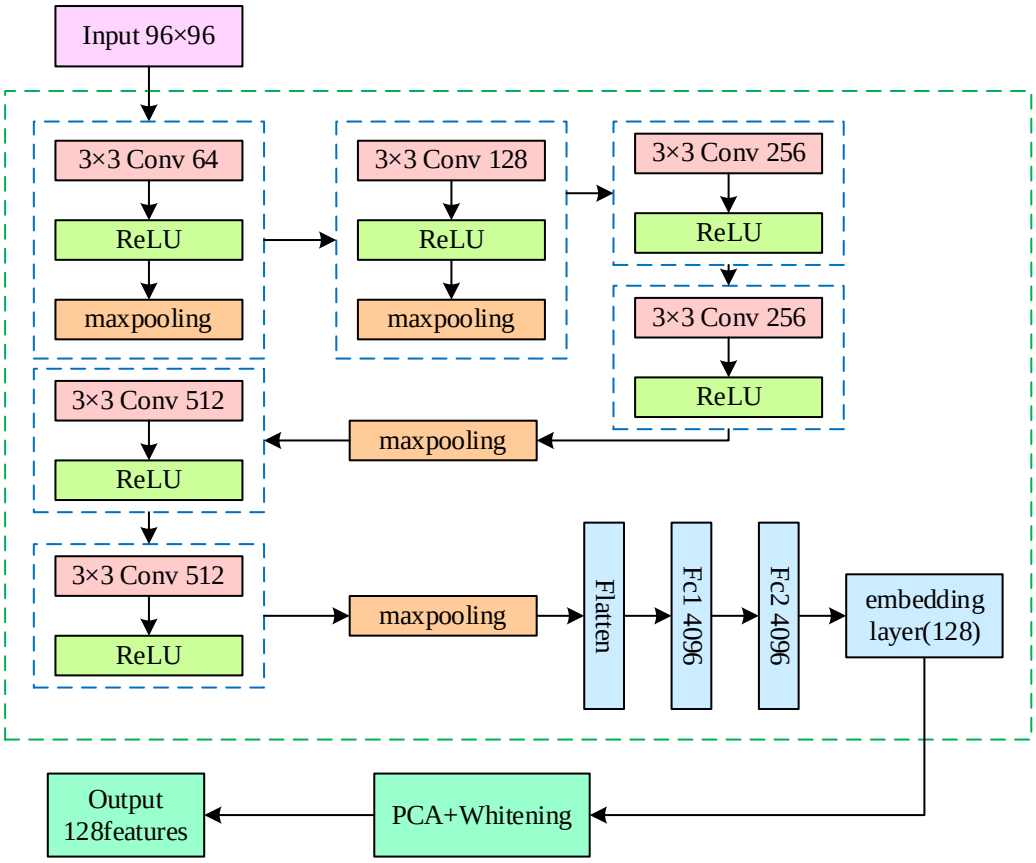


Figure 3: VGGish model and its later processing

The model is constructed by converting the initial data into MFCC features through the short-time Fourier transform to get the spectrogram, every 0.96s for a frame of data 96*64 such a feature image as input, the first layer uses 64 3*3 feature detectors, using the Relu activation function and added a pooling layer, to get the depth of 64 feature maps, the more feature detectors are used the more features are extracted, the second layer of the convolutional network uses 128 feature detectors, still performing Relu operations, and finally also added a pooling layer to get 128 feature maps, the next two layers use 265 feature detectors. The more feature detectors are used, the more features are extracted, the second layer of convolutional network uses 128 feature detectors, and still performs Relu operation, and finally adds a pooling layer to get 128 feature maps, the next two layers use 265 3*3 feature detectors, and a pooling layer is added after these two layers, the next two layers use 512 3*3 feature detectors, and the same as in the first two layers, a pooling layer is added to

the two layers convolution, and 512 feature maps are obtained, the next two layers use 512 3*3 feature detectors, and the same as in the previous two layers, a pooling layer is added to get 512 feature maps. The next two layers use 512 3*3 feature detectors, add a pooling layer with the same two-layer convolution as the previous two layers, and get 512 feature maps, and finally add three fully-connected layers, the fully-connected layer is the final output layer, and the fully-connected layer outputs a 128-dimensional feature embedding in this paper. The later data pre-processing, remove the head and tail of each 6 frames, 0.96s for a frame, try to 15 frames, 18 frames, 30 frames for a splicing for processing.

3.2 Data downscaling visualization

3.2.1 Dimensionality reduction methods

The dimensionality reduction method used in this paper, t-SNE, is the most effective data dimensionality reduction and visualization method, and its drawbacks are also obvious, such as: large memory occupation, long running time. However, in the initial determination of the divisibility of the data the advantages of this method are still greater than the disadvantages, the uncertainty of high-dimensional data through dimensionality reduction projected to the 2-dimensional or 3-dimensional data visualization, in the specific image to observe whether the interval between the same class is small, but the interval between different classes is large. Because of the loss of some information compared with the original data after the dimensionality reduction, if the data is separable in the low-dimensional space, the dataset is separable; if it is not separable in the low-dimensional space, it may be that the data itself is not separable, or it may be that the data can't be projected to the low-dimensional space. t-SNE is evolved from the SNE algorithm, which maps the data points to the probability distribution by affine transformation, and it consists of two main steps:

(1) SNE constructs a probability distribution between high-dimensional objects, so that similar objects have a higher probability of being selected and a lower probability of selecting dissimilar objects;

(2) SNE in the low-dimensional space in the construction of these points of the probability distribution, so that the two probability distributions between as similar as possible.

The method used by SNE to express point-to-point similarity is to transform the Euclidean distance into conditional probability, i.e., for the given N high-dimensional data $x_1,\cdots,x_N$, SNE firstly computes the conditional probability p_{ij}, which expresses the similarity between x_i and x_j, i.e., equation (10):

$$p_{j|i}=\frac{\exp\left(-\left\|x_i-x_j\right\|^2\Big/\left(2\sigma_i^2\right)\right)}{\sum_{k\neq i}\exp\left(-\left\|x_i-x_k\right\|^2\Big/\left(2\sigma_i^2\right)\right)} \tag{10}$$

For different x_i parameter δ_i takes different values, since the problem is now to solve the similarity between different two points, make $p_{ik}=0$. And for y_i in low dimensions specify the variance of the Gaussian distribution as $\frac{1}{\sqrt{2}}$, the similarity between two points of low dimensional data is calculated as in equation (11):

$$q_{j|i} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k \neq i} \exp\left(-\left\|y_i - y_k\right\|^2\right)} \tag{11}$$

Similarly let $q_{i|y} = 0$.

If the dimensionality reduction is good and the local features are well preserved, then $p_{j|i} = q_{j|i}$, so only the distance between the two distributions needs to be optimized-KL dispersion, KL dispersion is a measure of the asymmetry of the difference between two probability distributions P and Q, and the objective function is shown in Equation (12):

$$C = \sum_i KL\left(P_i \| Q_i\right) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{12}$$

The P_i represents the conditional probability distribution of all other data points given the point X_i. When applying the KL dispersion, it is important to note that the KL dispersion is not symmetric, so the penalty coefficients corresponding to different distances in the low-dimensional mapping are different, i.e., if the two closer points are used to simulate the two farther points, a smaller cost will be generated, and on the contrary, using the two farther points to simulate the two closer points will generate a larger cost, i.e., the SNE is more tends to preserve localized features in the data.

Since different points take different σ_i methods, the entropy of P_i firstly increases with the increase of σ_i. In this regard, SNE uses binary search to find σ, i.e., the concept of optimal perplexity is adopted, and the perplexity is shown in equation (13):

$$Perp(P_i) = 2^{H(P_i)} \tag{13}$$

$H(P_i)$ is the entropy of P_i, i.e., equation (14):

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i} \tag{14}$$

The number of effective nearest neighbor points near a point can be taken as an explanation for the perplexity. the SNE is relatively robust to adjustments in perplexity, usually chosen to be between 5 and 50, and given that, a bisection search is used to find the best σ.

The central question now becomes how to solve for the gradient, the objective function is equivalent to $\sum\sum -p\log(q)$ This form is similar to softmax, however, the objective function for softmax is $\sum -y\log p$, which corresponds to a gradient of $y - p$. Analogous to the gradient of softmax, we can derive the gradient of the conditional probability of i under j in the objective function of SNE as $2(p_{j|i} - q_{j|i})(y_i - y_j)$, and similarly the gradient of the conditional probability of j under i as $2(p_{j|i} - q_{j|i})(y_i - y_j)$. Finally, the complete gradient is obtained as Eq. get the complete gradient as in equation (15):

$$\frac{\delta C}{\delta y_i} = 2\sum_j \left(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j} \right)\left(y_i - y_j \right) \tag{15}$$

For the initial σ selection, a Gaussian distribution under a smaller σ is used for initialization. In order to speed up the optimization and avoid falling into a local optimum, the exponential decay term of the previous gradient accumulation is introduced in the update, which is a relatively large momentum to be used in the gradient as in Eq. (16):

$$Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta Y} + \alpha(t)\left(Y^{(t-1)} - Y^{(t-2)} \right) \tag{16}$$

Y_t is the solution at t iterations, η is the learning rate, and α_t is the momentum at t iterations.

In order to avoid local optimal solutions, an appropriate amount of Gaussian noise is introduced in each iteration during the initialization phase, and the noise is gradually reduced afterwards.SNE needs to run several optimizations in order to determine the hyperparameters such as the amount of the selected Gaussian noise, the time to start the decay, the learning rate, and the choice of the momentum to be completed. To solve the crowding problem, the t-SNE method is introduced, which differs from the traditional SNE in the following three points:

(1) Using a symmetric version of SNE to simplify the gradient formulation;

(2) Using t distribution instead of Gaussian distribution to express the similarity between two points in low-dimensional space.

(3) To avoid crowding, a t distribution with longer tails is used under low-dimensional space.

3.2.2 Symmetric SNE

The conditional probability distribution is replaced by a joint probability distribution, i.e., P now refers to the joint probability distribution of the points in the high-dimensional space, and Q is the joint probability distribution of the points under the low-dimensional space, and the objective function is Eq. (17):

$$C = KL\left(P \| Q \right) = \sum_i \sum_j p_{i,j} \log \frac{p_{ij}}{q_{ij}} \tag{17}$$

p_{ii}, q_{jj} is 0. Assume that $q_{ij} = q_{ji}$ for any $p_{ij} = p_{ji}$, and by this assumption call this type of SNE as symmetric SNE (symmetric SNE), and the probability distribution is rewritten as equation (18):

$$p_{ij} = \frac{\exp\left(-\left\| x_i - x_j \right\|^2 / 2\sigma^2 \right)}{\sum_{k \neq i} \exp\left(-\left\| x_k - x_i \right\|^2 / 2\sigma^2 \right)}, q_{ij} = \frac{\exp\left(-\left\| y_i - y_j \right\|^2 \right)}{\sum_{k \neq i} \exp\left(-\left\| y_k - y_i \right\|^2 \right)} \tag{18}$$

The use of a symmetric SNE expression, while succinct in its entirety, introduces the problem of outliers. To address this issue, the joint probability distribution definition is amended to Eq. (19):

$$p_{ij} = \frac{p_{ij|i} + p_{j|i}}{2} \tag{19}$$

It is guaranteed that $\sum_j p_{ij} > \frac{1}{2n}$, so that each point will have some contribution to cost. The symmetric SNE is shown in equation (20):

$$\frac{\delta C}{\delta y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j) \tag{20}$$

The congestion problem is also a shortcoming of the SNE algorithm, in which clusters are clustered together and cannot be distinguished without a clear boundary. In this regard, a Gaussian distribution is used in the high dimensional space to convert the distance into a probability distribution, and the t distribution, which is more favorable to the long tail and less susceptible to outliers, is used instead of the Gaussian distribution in the low dimensional space, so that the middle and low distances in high dimensions still have a large distance after mapping, and the similarity q through the transformations becomes Eq. (21):

$$q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k \neq i}\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}} \tag{21}$$

The t distribution is a superposition of infinitely many Gaussian distributions, which is computationally non-exponential and will be much easier. The optimized gradient is shown in equation (22):

$$\frac{\delta C}{\delta y_i} = 4\sum_j \left(p_{ij} - q_{ij}\right)\left(y_i - y_j\right)\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1} \tag{22}$$

3.3 Visualization of the value of dynamic design

3.3.1 Enrichment of musical information

Compared with the single sense of listening to music, visual dynamic design can convey more valuable music information through the combination of audio and visual. Due to the differences between ancient music and modern music aesthetics, contemporary people are not familiar with classical music scores, the existence of pipa scores and music scores is difficult for people without music background to listen to and appreciate, and in the case of modern music APP and music at hand, the tape version of the music scores seems to be very inconvenient, and the stage performances of the music scores, on the one hand, because of the stage performance of the existing stage performances have fixed locations, time and fixed cities, on the other hand, the background visual design of the stage performances is not well with the music culture and fixed cities. On the other hand, the background visual design of the stage performance is not well integrated with the music culture and feelings, and it is only the serial play of the music, and the cultural communication of the sheet music itself is not enough for the emotional resonance and value output of the modern audience. The dynamic visualization design can make the audience experience the music from listening and watching,

and even the offline tactile sensation with the help of audio-visual combination, so that they can quickly enter the cultural context and feel the music culture at that time, and for the people who live in a fast-paced life, this kind of visualization of the music presentation is more attractive and immersive.

3.3.2 Increased latitude of expression

Compared to the static music visualization lies more in the designer's personal artistic expression, if compared to the connotation of the musical work, the audience feels more is the artist's view of the music, and for the mobility of the music is not reflected, then the music of the sense of rhythm and rhythm will be sealed in the visual picture. The advantage of visual dynamic design lies firstly in the increase of time and space latitude from the visual latitude. At the same time, due to the popularization of electronic devices and the development of related technologies, the visualization of dynamic music can become more interesting and rich visual communication and integration of more interactive expression, for example, can be combined with some interactive operations in the mobile terminal, or can be intervened in the form of small games to pass the interactive, and in the future can be combined with the function of VR and AR, from the audio-visual as well as the interactive operation of the basis of adding a more real The tactile sensation can be added on the basis of audio-visual and interactive operation. Therefore, compared with traditional music appreciation and expression, dynamic visualization is a multi-dimensional music presentation with rich possibilities.

3.3.3 Broadening of dissemination channels

Compared to the popularity and spontaneity of current popular music, sheet music is not sufficiently disseminated. On the one hand, it lies in the limited form of its existence, and on the other hand, it is not electronic enough to adapt to the current communication media. With the popularization of smart mobile in the masses, people use more electronic devices for communication. So compared to the sheet music of the score, tape, stage performances, dynamic visualization of music can be MP3 format, can also be MP4 format, and even in the audio and video format category can be converted to each other, which makes the score can be played on various platforms, such as small programs, H5, web pages, pop-up windows, music software, video software, etc., and the traditional paper-based media have a very clear distinction between the visual and auditory Integration in a certain space so that the viewer in the convenient viewing at the same time has become more convenient, it can be said that the visualization of music dynamic design is more adaptable to the development of contemporary communication channels, and can also make use of the contemporary developed communication channels for a large number of publicity. At the same time, the research on music visualization can also be applied to the background stage of the performance of sheet music, and the background and related design can make the audience have a more immersive experience.

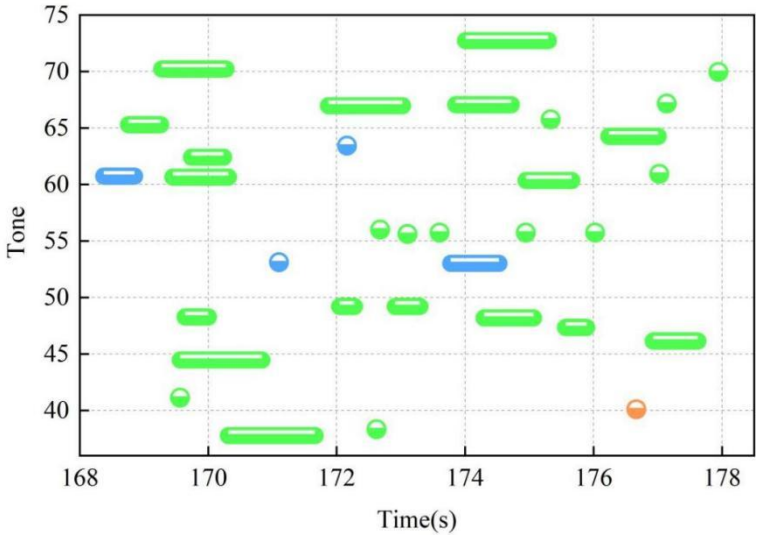
4 Application effect and development path of music visualization

This chapter examines, in turn, the transcription accuracy and overall performance of the multi-scale feature fusion-based score recognition model, the data processing of the VGGish model and the dimensionality reduction visualization effect of t-SNE, and the application effect of the proposed music data classification, dimensionality reduction and visualization

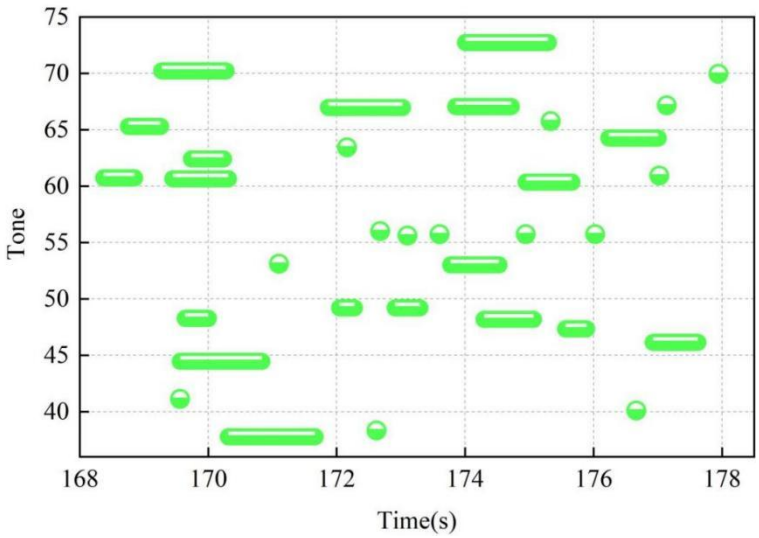
scheme. With the support of this AI tool, the development strategy of digital music visualization is discussed.

4.1 Effectiveness of the operation of the music score recognition model

The (X2) RGB video-based transcription algorithm and the (X1) score recognition model are selected to compare the transcription results of the two algorithms on the same piano performance piece in Fig. 4(a)-(b), in which the green frame is the normal transcription result, the blue frame is the missed detection, and the orange frame is the multi-detection, and the length of the frame corresponds to the pitch duration. The transcription algorithm based on RGB video has 4 missed detections and 1 multiple detection, while the score recognition model completes the transcription of all notes. It shows that the music score recognition model is able to fuse multi-scale data features, take advantage of features of different scales, and improve the transcription accuracy of the system.



(a) Transcription algorithm based on RGB video



(b) Textual method

Figure 4: Comparison of transcription results of the method

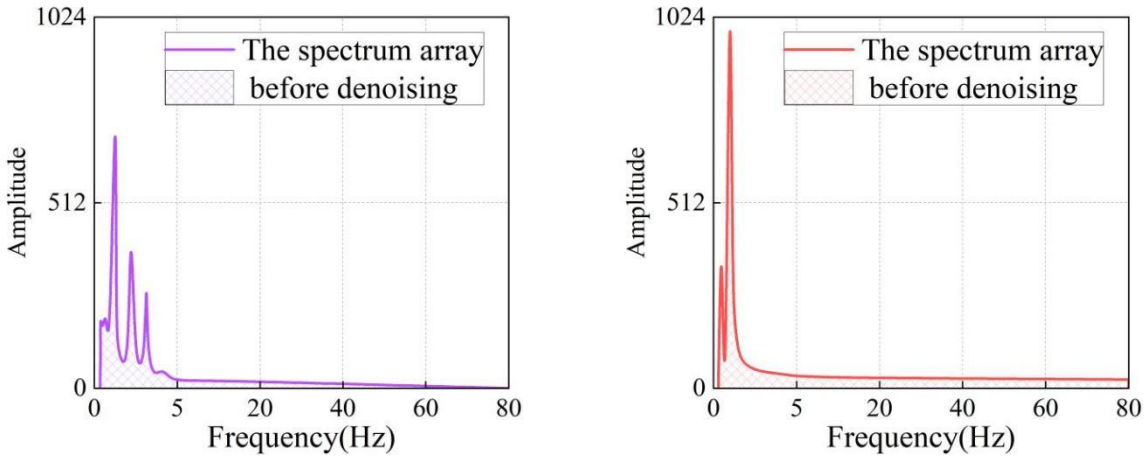
Further, (X3) Li, (X4) Koepke, (X5) RGB video based transcription module and (X6) skeleton based transcription module are added as control algorithms to enrich the comparison. The transcription accuracy, recall and F1 value performance on MIDI test set and OMP dataset are shown in Table 1. On MIDI test set dataset, (X1) music score recognition model has the best performance among algorithms in terms of accuracy (89.84%) and recall (89.95%), and the F1 value (85.15%) performs second only to (X4) Koepke algorithm (87.43%). On the OMP dataset, the data of the three evaluation metrics of the (X1) music score recognition model are the best among the six algorithms, and there is an accuracy rate with F1 value >90.00%, and the performance of the three metrics is 92.18%, 89.93%, and 95.55% in order. The effectiveness of the multi-scale feature fusion of the music score recognition model is further verified.

Table 1: The performance of the algorithm on the experimental dataset(%)

Algorithm	MIDI test set			OMP		
	Precision	Recall	F1	Precision	Recall	F1
X1	89.84	89.95	85.15	92.18	89.93	95.55
X2	77.57	79.61	63.03	81.79	82.47	89.84
X3	75.69	65.2	70.34	74.44	73.98	81.68
X4	84.82	87.64	87.43	80.04	77.13	70.1
X5	79.57	88.16	84.61	78.37	70.72	89.64
X6	82.36	66.4	60.54	83.03	80.08	86.02

4.2 Data Processing and Dimensionality Reduction Visualization

Playing any classical music (played with a monochord) on a computer and analyzing it using the VGGish model, the frequency and amplitude data in the spectrum array are visualized by lines in Fig. 5(a)-(b). By comparing the distribution of the lines at the same time period in the music playback, the corresponding frequency band of the monochord can be easily found. The sound of the dulcimer was found to be stable between 5 and 80 Hz in the 1024 Hz spectrum. Although the data in this frequency band will contain frequencies of some other noises, the music visualization requires less precision in the data and will not affect the subsequent parts of the design. The amplitude data in the range of 5 to 80 Hz will be considered as the frequency band of the monochord for subsequent data processing.



(a) Before denoising (b) After denoising

Figure 5: Comparison before and after denoising

The short-time fractal dimension reacts to the degree of difference between the amplitude transformations of the front and back frames of the music signal, which can react to the degree of intensity of the music transformation. t-SNE algorithm is used to visualize the fractal dimensions of each frame in Fig. 6. The fractal dimensions of each frame are in line with the actual performance of the classical music played, which preliminarily verifies the reliability of the t-SNE algorithm for the visualization of music.

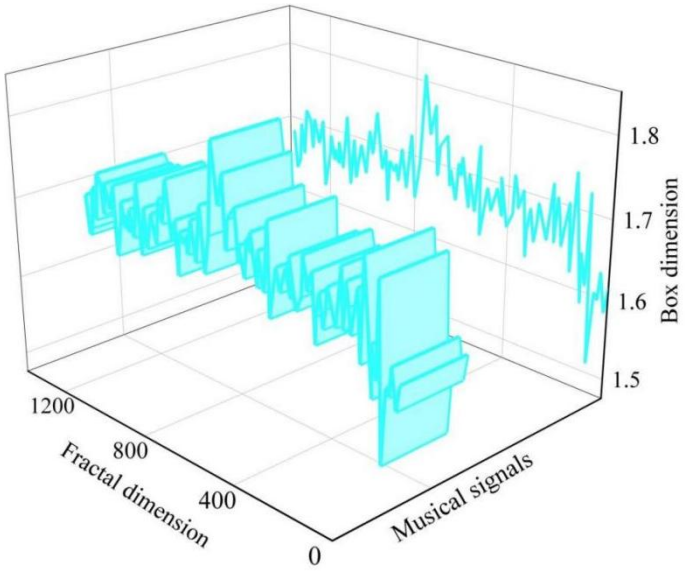


Figure 6: Short-term fractal dimension

The spectrogram reacts to the frequency characteristics of the music signal, and the output t-SNE algorithm visualizes the spectrogram of the classical music played in Fig. 7, which itself with color information can react to the music characteristics more intuitively, and the effect is better after visualization.

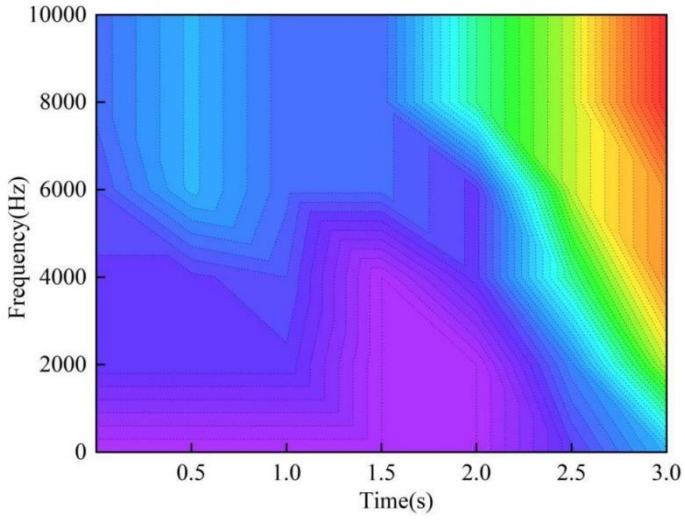


Figure 7: Spectrogram

4.3 Music Emotion Classification and Characteristic Importance

The musical score recognition model is used to form a feature matrix after audio feature extraction at the low level and music feature extraction at the middle and high levels of the

experimental music data, respectively. The feature matrix and emotion labels are divided into training set and test set in 3:1, and further classified using the music score recognition model to obtain the four classification results of music emotions and the feature ranking measured by feature contribution. The confusion matrix of the four emotions obtained from the classification is shown in Fig. 8. The music emotions are (Y1) Excitement, (Y2) Relaxation, (Y3) Irritation and (Y4) Depression, and the higher the predictive value between the emotion categories indicates a higher degree of identity. It can be seen that the predictive values of the model between the four categories of emotions and themselves are all 50.00 and above, and there are no cases where the predictive values of different categories of emotions are 50.00 and above. Overall the model's emotion classification accuracy is high, with categorization errors of 10.00% and below.

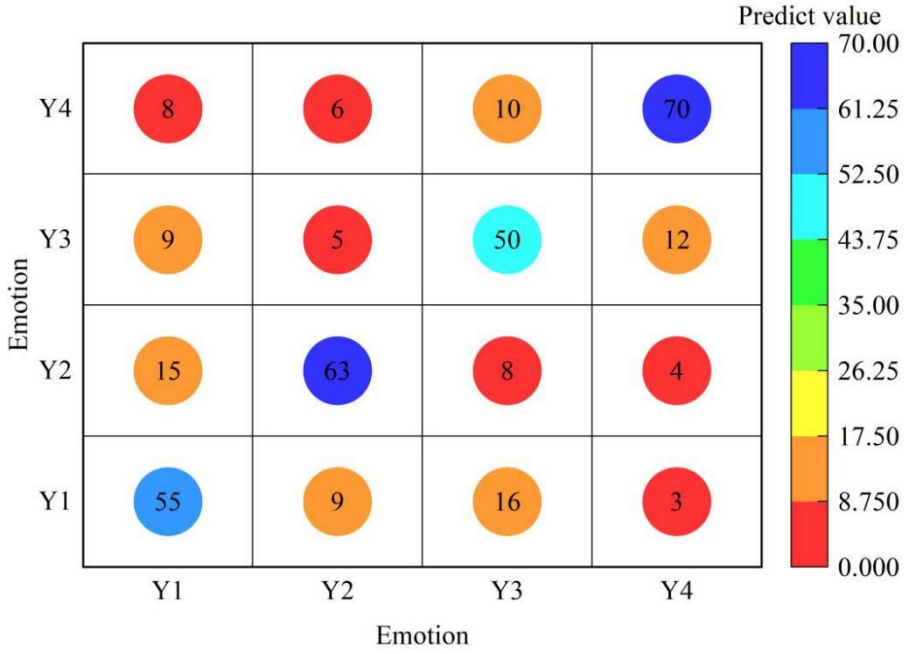


Figure 8: Confusion matrix for music emotion classification

The rankings of the four emotion classifications in terms of the importance of eight features, namely (Z1) average pitch, (Z2) spectral center of mass, (Z3) average tempo, (Z4) melodic alignment, (Z5) Mel's inverted spectral coefficient, (Z6) tempo, (Z7) over-zeroing rate, and (Z8) tonality, are shown in Fig. 9. The overall basic performances of the emotion classifications in terms of the importance of the features, from the highest to the lowest, are as follows: (Y2) relaxed, (Y4) depressed, (Y1) excited, and (Y3) annoyed.) depression, (Y1) excitement, and (Y3) irritability. The eight features mainly contributed more to (Y2) relaxation and (Y4) depression, and three of them, (Z1) average pitch, (Z2) spectral center of mass, and (Z3) average tempo, had an importance of 0.200 and above for (Y2) relaxation, which contributed more to relaxation, and could be used as a visual music feature for soothing mood.

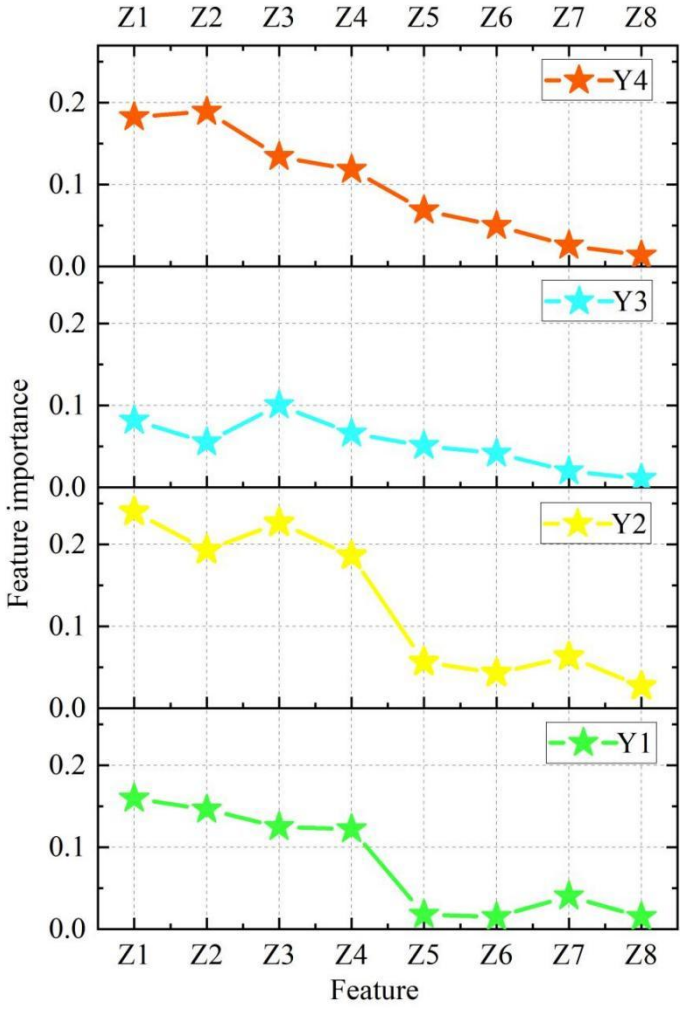


Figure 9: The ranking of feature importance in emotion classification

4.4 Overall application effect

300 subjects were selected as experimental samples, still using the classical music used above, using CRNN sequence classification network based on multi-scale feature fusion to classify and encode the optical score sequence of this music, and combining the VGGish model and t-SNE algorithm for the visualization of the music. Subjects' opinions were collected from two dimensions: music performance effect and visualization effect, in which the specific questions and options of the two dimensions are as follows:

(1) Are you satisfied with the effect of music performance of this playback:(A) Very satisfied, (B) Satisfied, (C) No feeling, (D) Unsatisfied.

(2) Are you satisfied with this form of music visualization used in this experiment: (A) Very satisfied, (B) Satisfied, (C) Didn't feel it, (D) Dissatisfied.

The two dimensions of the 50 subjects' opinions are shown in Figure 10. Overall, more than half of the subjects held the attitude of "(B) Satisfied" in the two dimensions, whether it was the effect of the music performance or the visualization effect. Among them, 27.25% of the subjects were (A) very satisfied with the music performance effect, i.e., 80% of the subjects resonated with the music performance effect of the model-assisted output of this paper. As for the visualization effect, 15.03% of the subjects held the attitude of "(B) satisfied", and a total of 67.27% of the subjects agreed with the music visualization form of the model.

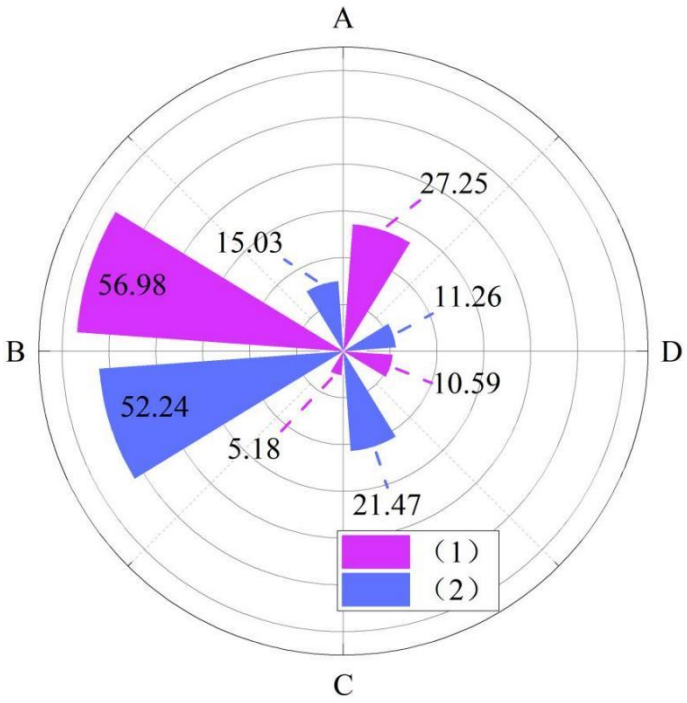


Figure 10: The opinions of the subjects in two dimensions

4.5 Suggested Development Strategies for Digital Music Visualization

The music score recognition model based on multi-scale feature fusion is used as a music score recognition and transcription tool, and the VGGish model and t-SNE algorithm are integrated as a music visualization method. Under this technological framework, the development path of digital music visualization in the AI era is proposed as follows:

(1) In the two-dimensional print media, it is necessary to cater to the characteristics and needs of the audience, and take dynamic media such as short video platforms as the main position for the creation, output and dissemination of digital music visualization. Actively expand the forms of creation and output, including but not limited to music videos, animated short films, short videos, microfilms and other forms, so as to give the audience a sense of freshness. Enriching the creation and output forms of digital music can not only promote the multi-form dissemination of musical works, but also stimulate the creators' creative energy and inspiration on the way of creation, thus promoting the innovation and development of the content and quality of musical works. In addition, in the selection of themes and contents of music works, it is also necessary to include the audience's preference and comprehension ability into important considerations, so as to promote the wide dissemination of digital music visualization works. In terms of the development strategy of digital music visualization in two-dimensional media, it is necessary to clearly define the audience of its own works, to target its preferences, and to focus on the audience's tastes and platform preferences for the creation, output and even dissemination of digital music visualization works.

(2) In addition to two-dimensional plane media, digital music visualization works can also use three-dimensional stereoscopic images as a carrier to realize their own creative innovation and dissemination development. In the three-dimensional image of the field performance stage, it should be matched with the type and content of the stage performance, and output digital music visualization works corresponding to it, thus stimulating the audience's musical perception and emotional feelings, and assisting in releasing the artistic expressiveness of the stage performance while completing the output of its own works and expression of

connotation. The combination of audio-visual music visualization can shape the emotional context that matches with the music works, making the audience produce strong emotional resonance, thus realizing the high-quality development of digital music visualization.

5 Conclusion

In this paper, we propose a music score recognition model containing four modules, namely, note sampling, note contour refinement, note classification prediction and sample learning. The model has no over-checking or under-checking in music score transcription comparison experiments, and has the highest precision, recall and F1 value of 92.18%, 89.95% and 95.55% in the experimental dataset in that order, and the categorization error degree of the four types of emotions is 10.00% and below . And use VGGish model to process music data, t-SNE algorithm for music data visualization, the output results of the collaboration between the two can intuitively reflect the musical characteristics of the experimental samples. The practical application works combining the three techniques, the music performance effect and the visualization effect are recognized by more than 80.00% and 50.00% of the subjects respectively. Relying on the proposed technology, it is suggested that digital music visualization should take two-dimensional print media and three-dimensional stereoscopic images as the main development direction, and combine the characteristics of different platforms to carry out the cross-cultural creation and output of traditional music art.

About the Author

Yujia Yang was born in Luoyang, Henan Province, China. She graduated from Tianjin Conservatory of Music, where she earned her Master’s degree. She is currently pursuing her Ph.D. at Universiti Putra Malaysia. Her primary research interests focus on music education and music psychology.

Dan Shen was born in 1980, in Hengyang, Hunan Province，P.R. China. She studied in Belarusian State Academy of Music and received her Master's degrees in both piano performance and chamber music performance in 2004. She is a dedicated lecturer, working at School of Art, South China University of Technology. Her research interests include Music Education and Piano Performance.

Corresponding Author. E-mail: shendan2025@163.com

Xuandong Sun was born in 1972, in Jining, Shandong Province,P.R. China. He obtained his doctoral degree from Chengdu Institute of Computer Applications, Chinese Academy of Sciences in 2012. He is currently working at the School of Computer Science, Guangdong University of Technology. His research direction is computer information technology and graphics processing.

References

[1] Ning, H., & Maneewattana, C. (2024). The characteristics and forms of contemporary Chinese Zheng music composition. Journal of Roi Kaensarn Academi, 9(3), 405-419.

[2] Yuan, Z. (2025). The Relationship between Chinese Music History and The Study of Traditional Chinese Music. Mediterranean Archaeology & Archaeometry, 25(1).

[3] Lee, D. (2025). Organising music’s structures: The classification of musical forms in

Western art music. Journal of Information Science, 51(3), 705-719.

[4] Posikura-Omelchuk, N., Ivannikov, T. Y. M. U. R., Molchko, U. L. Y. A. N. A., Prokopchuk, V. I. K. T. O. R. I. I. A., & Tatarnikova, A. N. G. E. L. I. K. A. (2025). The evolution of national performance art: musical interpretation in the context of tradition and modernity. Studia Universitatis Babes-Bolyai Musica, 70, 63-79.

[5] Jiayang Li, D. F. A., & Su, Y. (2024). Exploring the significance of traditional music in safeguarding and transmitting intangible cultural heritage: A case study of the Yunnan Bai ethnic group. Cultura: International Journal of Philosophy of Culture and Axiology, 21(3).

[6] Huang, Y., Chuangprakhon, S., & Santaveesuk, P. (2024). Preservation and transmission of Shaanxi Guzheng musical instruments: Challenges and strategies for cultural sustainability. International Research Journal of Multidisciplinary Scope, 5(04), 147-158.

[7] Gong, Y., Jirajarupat, P., & Zhang, Y. (2024). Guidelines for Literacy Transmission and Preservation of Bayu Folk Songs. International Journal of Education and Literacy Studies, 12(2), 94-100.

[8] Wu, W. (2023). Traditional Music Education Content and Environment's Influence on the Preservation of Traditional Music in Henan Province. Frontiers in Art Research, 5(17), 60-66.

[9] Cao, Y., & Park, J. (2022). Research on Visual Design of Traditional Music Based on AI Enabling Guided by Intangible Cultural Heritage Inheritance Concept. Frontiers in Art Research, 4(17), 32-35.

[10] Chen, D., Sun, N., Lee, J. H., Zou, C., & Jeon, W. S. (2024). Digital technology in cultural heritage: construction and evaluation methods of AI-Based ethnic music dataset. Applied Sciences, 14(23), 10811.

[11] Bosi, M., Canazza, S., Pretto, N., Russo, A., & Spanio, M. (2024). From Tape to Code: An international AI-based standard for audio cultural heritage preservation Don’t play that song for me (if it’s not preserved with ARP!). IEEE Access.

[12] Kuremoto, T. (2024). Guqing music recognition by machine learning methods. Impact, 2024(1), 40-42.

[13] 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.

[14] Chinnasamy, R. K., Saravanan, N., Gopalswamy, N., & Kumar, P. R. (2025, April). Music lyrics generator and translator. In AIP Conference Proceedings (Vol. 3279, No. 1, p. 020102). AIP Publishing LLC.

[15] Wei, H. (2025). Intelligent Music Recommendation System Using Fuzzy Convolutional Generative Adversarial Network of Personalized Music Experience. International Journal of High Speed Electronics and Systems, 2540188.

[16] Bi, Y. (2025). Vibrating Outside Borders Studying the Universal Outreach of Chinese Music Arts through Music Interaction. Cultura: International Journal of Philosophy of Culture and Axiology, 22(4).

[17] Eraslan, I. (2025). Artificial Intelligence and the Rewriting of Musical Memory: A Cognitive Perspective. Science Development, 6(3), 114-120.