Classification of Ethnic Music Styles Combining Timbre Clustering and Rhythmic Pattern Matching

Zhen Zhao
School of Arts, Guizhou University of Engineering Science, Bijie 551700, Guizhou, China
Corresponding author: Zhen Zhao (e-mail: ggc2025@163.com)

ABSTRACT Current research on ethnic music style classification faces challenges related to timbre overlap in multi-instrument recordings and unstable representation of rhythmic structures, which reduce the reliability of style discrimination. This paper proposes a classification framework combining timbre clustering with rhythmic pattern matching. Robust spectral–temporal feature representation and sequential pattern analysis are increasingly important in intelligent acoustic sensing and electromagnetic signal processing applications, providing methodological support for complex signal interpretation across engineering domains. The proposed framework constructs independent yet complementary representations for timbre distribution and rhythm organization, enabling spectral and temporal ambiguities to be jointly constrained within a unified classification process. Spectral features are extracted using short-time Fourier transform (STFT), followed by the computation of 13-dimensional MFCCs, first-order differential coefficients, spectral centroids, roll-off rates, and spectral flux features. After principal component analysis, Gaussian mixture models are employed for timbre clustering to isolate instrument-related acoustic components. Rhythm features are subsequently generated from timbre-cluster-dependent energy envelopes and aligned with ethnic rhythm templates through Dynamic Time Warping, allowing robust similarity measurement under nonlinear temporal variations. The timbre cluster probability distribution and rhythm matching descriptors are fused into a multidimensional feature vector and classified using a support vector machine. Experimental results demonstrate that the proposed model achieves an overall accuracy of 90.9% and an average F1 score of 91.3% for six ethnic music styles, validating the effectiveness and stability of jointly modeling timbre clustering and rhythmic pattern matching. The proposed framework provides a reliable strategy for complex spectral–temporal signal analysis and offers potential methodological insights for advanced electromagnetic sensing and intelligent signal interpretation.

INDEX TERMS timbre clustering, rhythmic pattern matching, gaussian mixture model, spectral-temporal signal analysis, electromagnetic signal processing

I. INTRODUCTION

ETHNIC music carries the core characteristics of regional culture and ethnic identity, and its style classification is of great significance for the digital preservation of music, cultural dissemination, and intelligent recommendation. The analysis is based on a curated dataset of multi instrument ethnic music recordings reflecting distinct stylistic categories. However, ethnic music presents a mixture of multiple instruments in terms of timbre structure, and has non-uniform and weak periodic characteristics in terms of rhythm. Traditional methods are difficult to accurately analyze its complex acoustic features [1,2]. The interaction between timbre and rhythm forms a key dimension in identifying ethnic music styles, while timbre overlap and rhythm variability introduce independent sources of ambiguity during feature representation [3,4]. Building a feature modeling framework that considers the correlation between timbre levels and rhythm patterns has become a core direction for improving the accuracy and interpretability of ethnic music style classification [5,6].

Previous studies have mainly attempted to solve the above problems from three aspects: deep feature modeling, rhythm structure analysis, and multimodal fusion. The spectral modeling method based on convolutional structure has made progress in capturing local timbre differences, but lacks the ability to express temporal evolution [7,8]. The rhythm pattern analysis model based on recurrent networks can enhance the temporal consistency of rhythm, but fails to distinguish the rhythm contributions between multiple sources [9,10]. There is still a problem of feature decoupling between timbre modeling and rhythm analysis in existing research, and there is a lack of cross layer correlation

mechanisms that can simultaneously constrain sound source distribution and rhythm structure, resulting in insufficient consistency in the representation of multi-source rhythm interaction in the model. This insufficiency arises from the lack of conditional dependency between sound source composition and rhythmic evolution, where rhythm descriptors are commonly derived from globally aggregated acoustic signals. As a result, spectral and temporal characteristics are often modeled in isolation, limiting the consistency of style representation across complex musical textures. In response to this structural deficiency, this article takes the collaborative modeling of acoustic statistical features and temporal patterns as the core idea, and constructs a unified representation framework for timbre distribution and rhythm patterns to achieve coordinated mapping of style features in the spectral and temporal domains. Model performance is assessed through supervised classification with training and testing partitions fixed by a consistent cross validation scheme. The periodic features are calculated through the rhythm graph and dynamically matched with the ethnic rhythm template to obtain rhythm similarity. The weight of timbre cluster distribution is combined with rhythm matching feature to construct a high-dimensional feature vector as input of SVM classification model to finish the style recognition. This paper carries out structured clustering in timbre level, constructs temporal template matching in rhythm level, and builds a timbre-rhythm collaborative representation framework with dual feature fusion, and provides a new methodological framework and theoretical reference for ethnic music style classification.

II. DESIGN OF STYLE CLASSIFICATION METHOD

Its overall architecture is inspired by the idea of hierarchical modeling, and the recognition of ethnic music style is realized by multi-level acoustic feature extraction and integration. The whole method framework is divided into three layers: input feature layer, representation modeling layer, and fusion classification layer, which reflect the hierarchical mapping relationship from low-level signals to high-level semantics. The input layer is used to extract and reduce the time-frequency features of the original audio, and the middle layer is further divided into two modeling paths: timbre and rhythm. The output layer completes feature fusion and classification discrimination. This hierarchical structure enforces dependency across acoustic layers, where timbre clustering constrains the temporal domain on which rhythmic organization is subsequently characterized, and establishes a reasonable balance between acoustic complexity and style abstraction in the model structure.

As shown in Figure 1, the input audio signal firstly implements STFT to get the spectral information, and then implements MFCC and spectral features extraction. After PCA dimensionality reduction, it enters the representation modeling layer. The cluster probability distribution output by the timbre clustering part and the descriptors extracted from rhythm features can be merged in the fusion part to form a multidimensional joint feature vector, and then classify the style by SVM. This hierarchical framework reflects the whole process of ethnic music style judgment from time-frequency analysis to high-level semantics, and provides overall logical support for further development of ethnic music feature extraction and classification algorithm.

AUDIO PREPROCESSING AND FEATURE EXTRACTION

The input signal is a multi-instrument mixed ethnic music waveform sequence of length L, denoted as $x(t)$, with a uniform sampling frequency of $f_s = 44.1$ kHz. After normalization, STFT is performed [11,12]. Hamming window $w(n)$, window length $N = 2048$, frame shift $H = 512$ are added to each frame, and the complex spectrum is calculated:

$$X(k,m) = \sum_{n=0}^{N-1} x(n+mH)w(n)e^{-j2\pi kn/N} \quad (1)$$

In formula (1), k is the frequency index, and m is the frame index. The square of the modulus is taken to obtain the power spectrum $P(k,m) = |X(k,m)|^2$, and a 40 V Mel filter group $M_r(k)$ is applied to obtain the filter energy output $E_r(m) = \sum_k P(k,m)M_r(k)$. The logarithm of $E_r(m)$ is taken, and the discrete cosine transform is performed. The first 13 dimensions of MFCC $c_p(m)$ and its first-order difference are taken to form a 26 dimensional MFCC feature matrix C_m. The spectral centroid $S_c(m) = \sum_k kP(k,m)/\sum_k P(k,m)$ and spectral roll off rate $S_r(m)$ are simultaneously calculated as the frequency position at which the cumulative spectral energy reaches 85% of the total energy, and spectral current $S_f(m) = \sqrt{\sum_k (P(k,m) - P(k,m-1))^2}$. Additional 3D spectral features (spectral centroid, spectral roll off rate, spectral stream) are extracted and concatenated to obtain a 29 dimensional timbre feature matrix.

A high-dimensional timbre feature matrix $F = [C_m, S_c, S_r, S_f] \in \mathbb{R}^{T\times d}$ is formed by concatenating the above three types of features, where T is the frame rate, and d is the feature dimension. To reduce redundancy between features and enhance separability, PCA is performed after standardizing F. This transformation is applied to decorrelate heterogeneous spectral descriptors and to align the feature space with the dominant variance directions governing timbral distribution. The covariance matrix is calculated:

$$\Sigma = \frac{1}{T-1}(F-\bar{F})^\top (F-\bar{F}) \quad (2)$$

The covariance structure reflects joint variability across cepstral and spectral dimensions, providing a global measure of inter-feature dependency prior to clustering. Feature decomposition $\Sigma v_i = \lambda_i v_i$ is performed on equation (2), and the eigenvectors corresponding to the first q largest eigenvalues are taken to form a projection matrix V_q, and obtain the dimensionality reduction result:

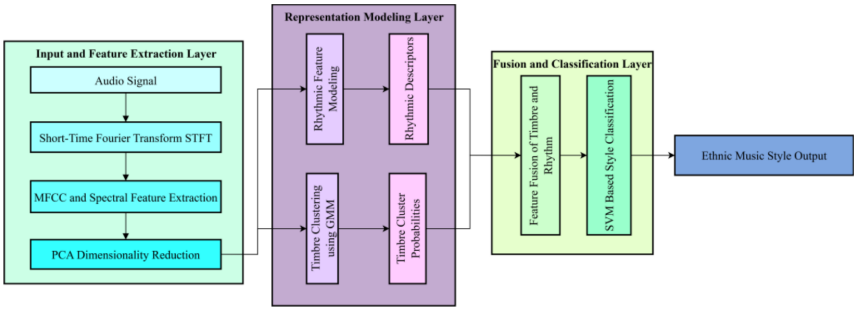

FIGURE 1. Overall Architecture of Ethnic Music Style Classification Method

$$Y = FV_q \tag{3}$$

The retained components correspond to the principal subspace preserving the dominant energy distribution of the original feature set while suppressing low-variance dimensions associated with noise and feature collinearity. In equation (3), $Y \in \mathbb{R}^{T\times q}$ is the reduced dimensional timbre feature matrix, and q is determined by the point at which the principal subspace stabilizes the variance structure required for probabilistic modeling of timbre distributions. The dimensionality reduced features significantly reduce the collinearity of spectral information while maintaining timbre differences, improving the discriminative efficiency of subsequent GMM clustering. This reduced representation preserves inter-feature contrast essential for Gaussian component separation, ensuring that the probabilistic clustering process remains sensitive to distinct timbral patterns.

TONE CLUSTERING

To achieve timbre separation of multiple sound source signals, GMM [13,14] is used for unsupervised clustering modeling of timbre distribution in the reduced dimensional space. Assuming that the timbre features are generated by K hidden sound sources, their probability density is represented as:

$$p(y_t|\Theta) = \sum_{k=1}^{K} \pi_k N(y_t|\mu_k, \Sigma_k) \tag{4}$$

The clustering is performed in the principal component space to balance model compactness and distributional fidelity, avoiding overfitting caused by redundant dimensions while preserving separability among sound source patterns. In formula (4), y_t is the feature vector of the tth frame; π_k is the mixed weight of the kth Gaussian component; μ_k and Σ_k are the mean vector and covariance matrix, K respectively; the parameter set is denoted as $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$. To obtain the optimal parameters, the expected maximization algorithm is used for iterative updates, and the posterior probability is calculated at the expected step:

$$\gamma_{tk} = \frac{\pi_k N(y_t|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(y_t|\mu_j, \Sigma_j)} \tag{5}$$

The parameters in the maximization step are updated as:

$$\pi_k^{new} = \frac{1}{T}\sum_t \gamma_{tk}, \quad \mu_k^{new} = \frac{\sum_t \gamma_{tk} y_t}{\sum_t \gamma_{tk}},$$
$$\Sigma_k^{new} = \frac{\sum_t \gamma_{tk}(y_t - \mu_k^{new})(y_t - \mu_k^{new})^\top}{\sum_t \gamma_{tk}}$$

The convergence is determined by monotonically increasing the likelihood function $L(\Theta) = \sum_t \log p(y_t|\Theta)$. To avoid overfitting of the model, diagonal constraints are applied to the covariance matrix, and the optimal K^* is determined within the candidate cluster number $K \in [2, 12]$ using Bayesian information criteria [15,16]:

$$\mathrm{BIC} = -2L(\Theta) + K(q + q^2)\ln T \tag{6}$$

In formula (6), the first term reflects the model fitting degree, and the second term is the complexity penalty term. The model with the smallest BIC is selected as the final clustering structure. After clustering is completed, the cluster labels of each frame's timbre are determined by $\hat{k}_t = \arg\max_k \gamma_{tk}$, while the cluster probability distribution $\gamma_t = [\gamma_{t1}, \gamma_{t2}, \ldots, \gamma_{tK}]$ is retained to describe the uncertainty of timbre components, providing weight constraints for subsequent rhythm feature extraction. These probabilistic partitions define timbre-consistent temporal streams that serve as the structural basis for rhythm modeling.

EXTRACTION OF RHYTHM FEATURES FROM TONE CLUSTERS

The input is the frame level timbre cluster label sequence obtained through GMM clustering and the corresponding cluster probability distribution matrix $\Gamma = [\gamma_{tk}]^{T\times K}$, where T is the frame rate, and K is the number of timbre clusters. To characterize the energy changes and rhythmic organization of each timbre cluster on the time axis, a weighted energy envelope $e_k(t)$ is constructed for each cluster k, defined as:

$$e_k(t) = \sum_{m=1}^{T} \gamma_{mk} P(m)\delta(t - mH/f_s) \tag{7}$$

In formula (7), $P(m)$ is the power spectral energy of the mth frame; γ_{mk} is the posterior weight of the timbre cluster; H/f_s is the time interval corresponding to the

frame shift. This expression eliminates the energy error induced by tone mixing via probability weighting, so that the energy envelope represents the dynamic track of a single tone in time. This formulation embeds rhythmic variation within a fixed timbral condition, yielding a coupled temporal representation. All subsequent rhythm processing steps, including beat detection, beat stabilization, and rhythm map construction, are performed directly on this cluster-specific energy envelope, forming a self-contained rhythmic representation pipeline. To improve time-domain resolution, $e_k(t)$ is smoothed and standardized using a third-order Savitzky Golay filter [17,18]. The temporal variation characteristic of energy envelope of timbre clusters is the key to the generation of rhythmic features. Figure 2 is the whole process of extracting rhythm features from timbre clusters.

As shown in Figure 2, after smoothing the weighted envelope, the beat pulse is clear, and the rhythm graph matrix presents an energy cycle distribution along the time axis, indicating that the energy accumulation of rhythm events in different frequency bands has obvious regularity. This time-frequency structure provides stable input for subsequent rhythm template matching, enabling rhythm features to have consistent expression forms on time scales and energy levels. The beat rate bpm_{est} is obtained by identifying the dominant periodic peak in the normalized autocorrelation of the smoothed energy envelope, with peak selection constrained by envelope variance to adaptively suppress spurious periodicities. The energy envelope autocorrelation function is calculated:

$$R_k(\tau) = \sum_t e_k(t) e_k(t-\tau) \tag{8}$$

The maximum peak position τ_{max} within the beat search interval [0.3 s, 2.0 s] can be found, and the beat period $T_b = \tau_{max}$ and beat rate $bpm_{est} = 60/T_b$ are defined.

After taking the first derivative of the smooth envelope, the set of positive zero crossing points $\{t_i^k\}$ is taken to form a candidate set of starting points for the timbre cluster. A threshold is set to $\tau_b = 30/bpm_{est}$ (corresponding to half beat interval) based on the estimated beat period. When the interval between adjacent starting points is less than τ_b, only the points with higher energy amplitude are retained. This process stabilizes beat positioning in the time domain and enhances the significance of rhythm events. The rhythm map is defined as a beat-indexed and frequency-binned energy matrix derived from consecutive beat intervals, where each column corresponds to a stabilized beat segment and each row encodes rhythmic frequency energy. After completing the rhythm positioning, the local rhythm intensity between beats is calculated as:

$$r_i^k = \int_{t_i^k}^{t_{i+1}^k} e_k^2(t)\, dt$$

Using each starting point as a time anchor, a rhythm graph matrix on the entire sequence is constructed:

$$T_k(f,t) = \left| \sum_\tau e_k(\tau) e^{-j2\pi f\tau/T_b} \right| \tag{9}$$

The $T_k(f,t)$ in formula (9) represents the rhythmic energy distribution in the frequency f and time t dimensions, which is used to reveal the energy concentration of timbre clusters in different rhythmic frequency bands. To suppress amplitude differences, the rhythm chart is subjected to logarithmic normalization.

RHYTHM PATTERN MATCHING

To eliminate the influence of tempo variation and beat drift, dynamic time warping is employed to compute the optimal temporal alignment between rhythm feature sequences extracted from timbre-cluster energy envelopes and predefined ethnic rhythm templates [19,20]. Let the tone cluster rhythm vector sequence be $a = \{a_1, a_2, \ldots, a_M\}$; the ethnic rhythm template sequence be $b = \{b_1, b_2, \ldots, b_N\}$; the local cost function be defined as:

$$d(i,j) = \|a_i - b_j\|_2^2 \tag{10}$$

In formula (10), a_i and b_j respectively represent the rhythm feature vectors on the corresponding time frames, and the Euclidean distance is used to characterize the instantaneous rhythm differences [21]. In this formulation, both sequences represent time-ordered rhythm descriptors derived from beat-synchronous energy distributions, and no timbre descriptors are directly involved in the DTW alignment process. Temporal alignment is therefore performed within timbre-conditioned sequences, preserving the dependency between sound source structure and rhythmic similarity. A dynamic programming recursive relationship is constructed through the cumulative distance matrix $D \in \mathbb{R}^{M\times N}$:

$$\begin{aligned} D(i,j) = d(i,j) + \min[&D(i-1,j), D(i,j-1), \\ &D(i-1,j-1)] \end{aligned} \tag{11}$$

The boundary condition is $D(1,1) = d(1,1)$, and the normalized matching distance after recursion is defined as:

$$S_{k,s} = 1 - \frac{D(M,N)}{\max(M,N)\cdot d_{max}} \tag{12}$$

In formula (12), $S_{k,s} \in [0,1]$ represents the similarity between timbre cluster k and rhythm template s, and d_{max} is the upper bound of local distance. To improve adaptability to different levels of rhythm, the rhythm graph matrix $T_k(f,t)$ is subjected to multi-resolution convolution integration, smoothing the rhythm graph into a characteristic curve in the logarithmic frequency domain:

$$g_k(f) = \frac{1}{T}\sum_t T_k(f,t) \tag{13}$$

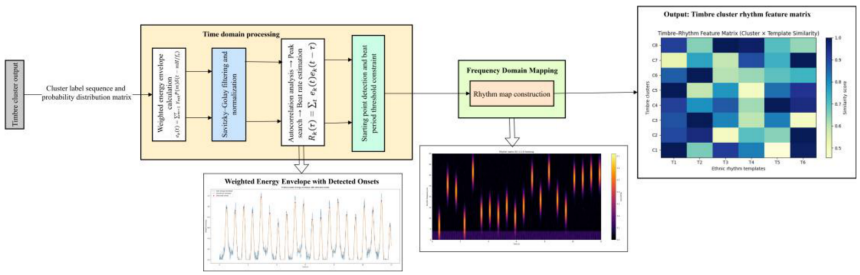

FIGURE 2. Schematic diagram of extracting rhythm features from timbre clusters

The $g_k(f)$ in formula (13) represents the rhythm frequency energy distribution function, and the frequency band misalignment caused by DTW path offset is compensated by comparing the correlation coefficient between $g_k(f)$ and template distribution $g_s(f)$. The final rhythm matching score is a weighted combination:

$$R_{k,s} = \alpha S_{k,s} + (1-\alpha)\rho(g_k, g_s) \tag{14}$$

In formula (14), $\rho(g_k, g_s)$ is the Pearson correlation coefficient, and α is the balance parameter between path matching and frequency domain similarity, which is determined through grid search on the training set.

After matching, the output matrix $R = [R_{k,s}] \in \mathbb{R}^{K\times S}$ represents the rhythmic similarity scores obtained by aligning cluster-specific rhythm feature sequences with ethnic rhythm templates, reflecting how rhythmic structures vary across timbre-conditioned temporal streams. This matrix provides a quantitative basis at the rhythm level for subsequent feature fusion and style classification, allowing the classification model to include both time aligned rhythm correlations in the input stage and preserve the rhythm difference structure between timbre clusters.

FEATURE FUSION AND CLASSIFICATION

To construct feature vectors with collaborative representation capabilities, the energy weighted averaging is first performed on each timbre cluster in the time dimension to form a timbre distribution vector $p = [p_1, p_2, \ldots, p_K]^\top$, defined as:

$$p_k = \frac{1}{N}\sum_{t=1}^{N} w_t P_{tk} \tag{15}$$

In formula (15), P_{tk} is the posterior probability that the tth frame belongs to the timbre cluster k, and w_t is the normalized weight of frame energy to enhance the contribution of high-energy spectral bands. This step integrates temporal timbre information at the probability level to generate a global timbre structure representation. To unify dimensions, p undergoes L2 standardization to ensure that each cluster weight has an equal scale measure in the feature space.

In the rhythm matching section, a rhythm feature vector $s_k = [s_{k1}, s_{k2}, \ldots, s_{kM}]^\top$ is constructed based on the similarity distribution between each timbre cluster and the rhythm templates of various ethnic groups, where s_{km} is derived from the matching score in formula (14). The rhythm vectors corresponding to all timbre clusters in cluster order are concatenated to form a rhythm feature matrix $R = [s_1^\top, s_2^\top, \ldots, s_K^\top]^\top \in \mathbb{R}^{K\times M}$. To avoid redundant coupling caused by high-dimensional features, singular value decomposition is performed on R: $R = U\Sigma V^\top$, and the left singular vectors corresponding to the first q singular values are taken to form a matrix U_q, obtaining the compressed rhythm feature:

$$r = U_q^\top R \tag{16}$$

In formula (16), $r \in \mathbb{R}^{q\times M}$ is the representation of redundant rhythm features, where q is determined by a cumulative energy contribution rate of 0.9. This step preserves the main rhythm variation patterns through low rank approximation and weakens the linear correlation between timbre clusters. To achieve the fusion of timbre and rhythm, a joint feature vector is defined:

$$f = [\alpha p^\top, \beta \mathrm{vec}(r)^\top]^\top \tag{17}$$

In formula (17), $\mathrm{vec}(r)$ represents matrix vectorization operation, and α and β are timbre and rhythm weight coefficients, respectively, satisfying $\alpha+\beta = 1$. The weight ratio is determined by minimizing the classification error on the training set through five fold cross validation. By integrating timbre statistics with rhythm similarity scores indexed through timbre-conditioned temporal streams, the proposed fusion constructs a structured joint representation in vector space, imposing simultaneous constraints on spectral structure and temporal patterns at the classifier input level. SVM adopts radial basis function $K(f_i, f_j) = \exp(-\gamma\|f_i - f_j\|^2)$, and the model training objective function is

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{T}\xi_i, \\ \text{s.t.} \quad & y_i(w^\top\phi(f_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0. \end{aligned}$$

w is the classification hyperplane weight vector; b is the bias term; ξ_i is the relaxation variable; C is the penalty factor; $\phi(\cdot)$ is the kernel mapping function [22,23]. Parameters γ and C are determined through grid search. The model output is a set of style labels $\hat{y} = \{y_1, y_2, \ldots, y_T\}$, where $y_i \in \{1, \ldots, L\}$ and L are the number of ethnic music styles.

III. EXPERIMENTAL DESIGN AND CONFIGURATION

DATASET AND PREPROCESSING

The data used in the experiment comes from a multi voice recording database of ethnic music, covering six styles of Han, Tibetan, Mongolian, Uyghur, Yi, and Korean ethnic groups. Each sample includes composite performance segments of string, wind, and percussion music, with a sampling rate of 44.1 kHz and a bit depth of 16 bits, consistent with the previous feature extraction process. The raw data is subjected to unified loudness correction and mute segment truncation, and after removing amplitude abnormal frames, it is stored in groups according to style. The length of each sample segment is limited in the range of $15 \sim 25$ s to preserve the rhythm structure. Then, to avoid the influence caused by different recording scenes, logarithmic amplitude normalization is applied in the spectral domain, and the short-term energy curve is smoothed by third order Savitzky Golay filter to keep the continuity of rhythm-peak structure. All samples are transferred to monaural to suppress the influence of phase difference on timbre clustering. Sample Partition Layered five fold cross is used to partition sample. Each fold contains a training set to test set, and the sample size of each style is balanced. The number of samples in the dataset is listed in Table 1. The coverage of each category of sample could be ensemble scene of different ethnic traditional musical instruments, and the generalization of clustering and rhythm matching algorithm under different sound source condition is guaranteed.

After obtaining the time–frequency representation via STFT, the signal undergoes logarithmic amplitude normalization within the Mel filter bank's energy domain, ensuring uniform energy distribution across heterogeneous recording conditions. The timbral feature matrix, subsequently reduced in dimensionality through PCA, is normalized at both frame and sample scales to preserve the numerical stability of the covariance estimation during GMM clustering.

PARAMETER SETTINGS

The determination of parameters for each stage of the experiment is based on the pre-processing results of the previous data and the performance requirements of the model. All parameters are fixed after cross validation and Bayesian information criterion optimization. The input for the timbre clustering stage is a 20 dimensional timbre feature matrix reduced by PCA, and the dimension is truncated when the cumulative contribution rate reaches 0.95. GMM is trained using the expectation maximization algorithm, and the number of clusters is fixed at 8 after searching between 2 and 12, corresponding to eight dominant timbre types. The covariance matrix is set to diagonal form to avoid parameter redundancy; the initial mixed weights are evenly distributed; the convergence threshold is set to 10^{-6}; the maximum number of iterations is 200. The model structure is determined based on the minimum value of Bayesian information criterion to maintain the balance and separability of clustering results. The rhythm feature extraction stage calculates the autocorrelation function on the energy envelope of the timbre cluster, with an autocorrelation search interval of 0.3 to 2.0 seconds and a step size of 0.01 seconds. The energy envelope is smoothed by Savitzky Golay filtering, with a window length of 9 frames and a polynomial order of 3. The beat detection threshold is dynamically adjusted based on the amplitude of the main peak of the signal, and the noise suppression factor is set to 0.15 to ensure stable beat intervals. The rhythm chart calculates a frequency range of 0.5 to 5 Hz (corresponding to a beat period of 0.2-2.0 s), with a frequency resolution of 0.05 Hz, and eliminates energy amplitude differences through logarithmic normalization. The path bandwidth of the dynamic time warping algorithm is set to 10% of the sample length; the cumulative distance upper limit is three times the average local distance; the path balance coefficient is set to 0.6 to maintain the ratio of time matching and frequency domain correlation.

In the feature fusion stage, the timbre and rhythm features are linearly combined in a weight ratio of 0.5:0.5, and the weight ratio is determined by minimizing the classification error through five fold cross validation. The rhythm features are compressed by singular value decomposition, and the dimensions are retained by extracting the cumulative energy contribution rate of 0.9. SVM adopts radial basis function, with penalty factor C searching exponentially within the range of [0.1,100], and kernel width γ adjusted within the range of $[10^{-3}, 10]$. The final parameter combination is obtained through grid search, which enables the model to converge and stabilize the error between the training and the test sets. The maximum iteration limit during the training phase is 1000, with a tolerance of 1×10^{-4}. All experiments use the same sample partitioning and feature input. The parameter system maintains consistent constraint relationships in the three stages of timbre clustering, rhythm matching, and classification, ensuring the repeatability of subsequent style classification and robustness analysis.

EVALUATION INDICATORS

The experimental evaluation focuses on the accuracy of style classification and consistency of feature structure, using multidimensional indicators to comprehensively measure the performance of timbre clustering, rhythm matching, and fusion models in ethnic music style recognition. The classification performance is mainly based on overall accuracy, which is used to reflect the recognition level of the model on six types of ethnic music samples. To measure the stability and balance of the prediction results, this paper calculates the macro average precision, macro average recall, and F1 score, which maintain equal weight evaluation of various styles under uneven sample size conditions. The F1 value is obtained by harmonic averaging of precision and recall, and is used to comprehensively describe the identification strength of the model for each category boundary. The separability of the fused feature space is indirectly verified through four indicators: inter class distance, intra class vari-

TABLE 1. Composition of Ethnic Music Dataset

Ethnic Category	Number of Samples	Average Duration (s)	Number of Parts	Sampling Rate (Hz)
Han	220	18.4	3	44100
Tibetan	200	19.2	3	44100
Mongolian	210	20.1	4	44100
Uyghur	215	17.9	3	44100
Yi	205	18.6	3	44100
Korean	210	19.5	4	44100

ance, Davies Bouldin (DB) index, and silhouette coefficient. The inter class distance reflects the degree of differentiation of different ethnic music styles in the fusion space, the intra class variance describes the concentration of samples in the same class, and the silhouette coefficient measures the overall clustering quality by the ratio of the two. To evaluate the adaptability of the model under multiple conditions, additional calculations are made to assess the changes in classification accuracy under different rhythm complexities and signal-to-noise ratios, in order to reflect the robustness of the fused features to interference. All indicators are calculated under a unified sample division, and the results are used for comparative analysis of the performance and stability of various methods in subsequent chapters.

IV. RESULT ANALYSIS

TONE CLUSTERING EFFECT

To evaluate the adaptability of different clustering algorithms in separating ethnic music timbres, a comparative clustering analysis was conducted on the spectral features of the same batch of multi instrument mixed audio samples. Method one is K-Means clustering, which implements hard partitioning based on the principle of minimizing Euclidean distance. The structure is simple but insensitive to complex timbre boundaries. Method 2 is Spectral Clustering, which identifies nonlinear boundary structures through similarity matrix feature decomposition and is suitable for smoothing spectral features. Method three is Density-Based Spatial Clustering of Applications with Noise (DBSCAN), which forms a clustering structure based on local density and is robust to transient signals. Method four is deep feature clustering using Variational Auto-Encoders (VAE)+K-Means, which maps the latent space using variational autoencoders before performing clustering to capture complex harmonic patterns. Method five is GMM, which fits continuous spectral features with probability distribution to achieve soft partitioning and reduce timbre overlap interference. The experiment divided the timbre features into eight clusters and speculated on the dominant timbre types of flute, pipa, guzheng, erhu, drum, xiao, suona, and gong, respectively, to compare the differences in timbre clustering purity (0-1) and intra cluster similarity (0-1) between different algorithms. The result is shown in Figure 3.

The clustering purity of GMM in Cluster 2, 4, and 6 reached 0.85, 0.88, and 0.89, respectively, and the intra cluster similarity was 0.88, 0.91, and 0.90, respectively, indicating that its probabilistic modeling has strong adaptability to the structure of continuous spectra such as string and orchestra. The purity of VAE+KMeans in Cluster 3 and 4 reached 0.87 and 0.85, with similarity of 0.91 and 0.88, reflecting the potential distinguishing force of deep features in high-dimensional harmonic modes. However, in Cluster 5 and 8, where transient features are significant, the purity decreased to 0.79 and 0.77, indicating insufficient sensitivity to short-term energy changes. The purity of spectral clustering in Cluster 1 and 6 reached 0.83 and 0.84, and the similarity was 0.88 and 0.87, indicating its good boundary recognition ability in smooth spectral boundaries of wind music timbres. The purity of DBSCAN in Cluster 5 and 8 was 0.84 and 0.83, and the similarity was 0.86 and 0.85, which was consistent with the high-energy transient structural characteristics of percussion tones. The purity of K-Means reached 0.83 in Cluster 5, with a similarity of 0.85, and it only performed outstandingly in timbre clusters with a single energy distribution.

RHYTHM MATCHING PERFORMANCE

To investigate the performance of rhythm alignment algorithm on different timbre clusters, this section conducted rhythm template matching analysis based on the results of eight timbre clustering. Timbre features were divided into eight clusters using a GMM, presumably corresponding to dominant timbre types such as flute, pipa, guzheng, erhu, drum, xiao, suona, and gong. Method 1 uses DTW to nonlinearly align rhythmic events, emphasizing local temporal elasticity; Method 2 uses frame-by-frame direct matching, calculating similarity at a fixed time step; Method 3 uses cross-correlation to measure the degree of rhythmic overlap using overall time delay; Method 4 utilizes spectral rhythm graph matching to measure the consistency of rhythmic energy distribution in the frequency domain. All four methods calculated rhythm matching correlation coefficients on the same timbre clusters, and used the Pearson correlation coefficient $r \in [-1, 1]$ to characterize the similarity between the rhythm template and the rhythmic sequence of the timbre cluster. Values closer to 1 indicate more consistent rhythmic structure. The results are shown in Figure 4.

The DTW method exhibited the highest and most stable matching correlation coefficient across all timbre clusters, with medians generally around 0.800, a concentrated bin range, and few outliers, demonstrating robust time alignment performance under complex rhythmic patterns. The Direct method's median was concentrated around 0.538 , with a wide range of fluctuations, reflecting significant rhyth-

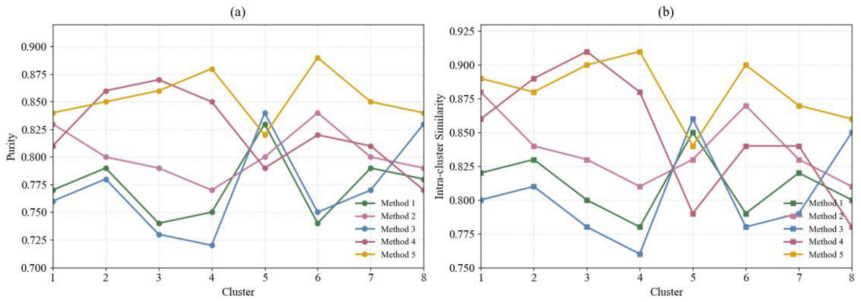

FIGURE 3. Comparison of clustering purity and intra cluster similarity of different clustering algorithms on various timbre clusters. (a)

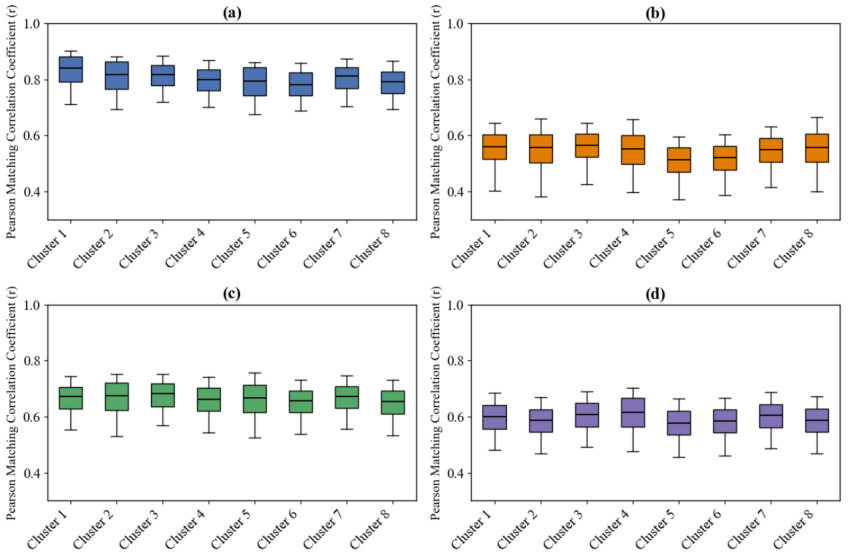

FIGURE 4. Comparison of correlation coefficient distribution for rhythm template matching. (a) DTW matching; (b) frame by frame direct comparison; (c) cross-correlation matching; (d) spectral rhythm map matching

mic misalignment due to its fixed time step nature. The CrossCorr method performed slightly better among clusters, with an overall median of approximately 0.660. Some clusters had higher upper quartile values due to the strong periodicity of their rhythmic structures. The median of the Spectral Matching method was about 0.592, with relatively stable fluctuations but overall low accuracy, indicating that although frequency domain matching can characterize the global rhythm energy distribution, it is difficult to handle local beat drift. The comprehensive results show that the nonlinear time alignment strategy has a matching advantage in the complex polyphonic rhythm structure of ethnic music.

SPATIAL DISTRIBUTION OF TIMBRE RHYTHM FUSION FEATURES

To investigate the performance differences between timbre features and rhythm features in distinguishing ethnic music styles, and to verify the enhancement effect of their fusion on feature separability, a two-dimensional feature space visualization map based on t-distributed stochastic neighborhood embedding (t-SNE) nonlinear dimensionality reduction was constructed. The horizontal and vertical axes respectively represent the two principal component directions of the sample in the low dimensional embedding space, reflecting the distribution relationship of high-dimensional features in the local neighborhood structure. (a) The subgraph displays the timbre space based on MFCC and spectral features, reflecting the ability of timbre attributes to distinguish styles; (b) The subgraph is a rhythm feature space composed of beat envelope and rhythm pattern matching features, used to reveal the clustering trend of rhythm structure; (c) The subgraph is the fusion space of timbre and rhythm features, combined with the probability distribution of timbre clusters and rhythm matching scores, reflecting the comprehensive role of joint features in style differentiation. The six colors correspond to music samples of Han, Tibetan, Mongolian, Uyghur, Yi, and Korean ethnic groups, respectively. The dot matrix in Figure 5 reflects the distribution structure of each category in the feature space.

The highly mixed distribution of sample points in the timbre feature space indicated that a single timbre attribute was insufficient to reflect stylistic differences, and some timbre overlap between styles blurred boundaries. The rhythm feature space showed a preliminary clustering trend, with ethnic groups with strong rhythmic regularity, such as Koreans and Uyghurs, forming relatively tight clusters, while samples from Tibetans and Mongolians with higher rhythmic freedom were more dispersed. In the fused feature space,

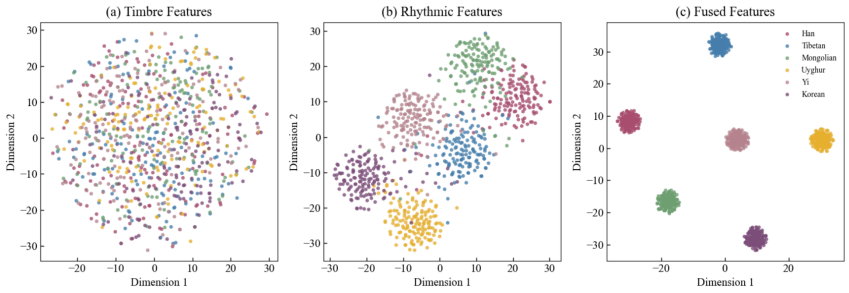

FIGURE 5. Visualization of Ethnic Music Style Feature Space. (a) Timbre Features, (b) Rhythmic Features, (c) Fused Features

clusters were completely separated with clear boundaries, indicating that the joint representation effectively enhanced inter-class discrimination. The complementary effects of timbre structure and rhythmic pattern were evident here, and the fusion strategy significantly improved the distinguishability of ethnic music styles. To further validate the clustering trends reflected in the visualization results, this paper quantitatively evaluated the separability of timbre, rhythm, and their fused feature spaces. Four metrics, including inter-class distance, intra-class variance, Davies–Bouldin (DB) index, and silhouette coefficient, were used to measure the clustering tightness and discriminability of the different feature representations. Table 2 showed that the inter-class distance in the timbre feature space was 1.857, and the intra-class variance was 3.412, indicating that the joint representation effectively enhanced inter-class discrimination. The clustering distribution was relatively loose, indicating significant overlap of timbre information in style representation. The inter class distance of the rhythm feature space increased to 3.941, and the intra class variance decreased to 2.093, indicating that the regularity of rhythm structure enhanced style differentiation. The inter class distance of the fused feature space increased to 7.926, and the intra class variance further decreased to 1.024. The DB index decreased to 0.615, and the silhouette coefficient increased to 0.708, indicating that the complementary relationship between timbre and rhythm formed a highly separable cluster structure after fusion. The joint representation of multi-source features achieves optimal discrimination in style discrimination, laying the foundation for the performance improvement of subsequent classification models.

STYLE CLASSIFICATION RESULTS

In the research of ethnic music style recognition, in order to verify the effectiveness of the fusion strategy of timbre clustering and rhythm pattern matching in classification performance, four different feature modeling methods were constructed for classification models, and a systematic comparison was conducted on six types of ethnic music styles - Han, Tibetan, Mongolian, Uyghur, Yi, and Korean. Model one is MFCC+SVM, which uses traditional MFCC to characterize timbre differences and evaluate the discriminative ability of basic timbre features; Model 2 is DTW+SVM, which is based on rhythm time series and template matching to examine the recognition contribution of rhythm dynamic structure; Model three is PCA+GMM+SVM, which introduces Gaussian mixture clustering and extracts cluster structure information in the reduced dimensional timbre space to reflect the distribution characteristics of instrument timbre; Model four is GMM+DTW+SVM, which integrates rhythm pattern similarity based on timbre clustering results to achieve multi-dimensional feature joint discrimination. Confusion matrix data comparisons across the six styles reveal the performance differences between the different feature modeling strategies in terms of timbre overlap, rhythmic coupling, and style discrimination. The results are shown in Figure 6.

The classification results of the fusion model showed significant advantages in overall accuracy and category differentiation. The average accuracy of MFCC+SVM was 76.7%, with a high confusion rate between Tibetan and Yi, indicating that single timbre features have limited discriminative power in multi-instrument signals. The average accuracy of DTW+SVM increased to 81.3%. The rhythmic patterns of Uyghur and Han peoples were more regular, and the beat characteristics enhanced the model's recognition accuracy of rhythmic structure. The average accuracy of PCA+GMM+SVM reached 84.1 %. Timbre clustering strengthened the timbre boundaries of instruments, significantly reducing misclassification between Tibetan and Mongolian. The fusion model GMM+DTW+SVM achieved an overall accuracy of 90.9%, with the main diagonal ratio exceeding 89 % for each category, and the most significant improvement in distinguishing Yi from Tibetan. To further verify the performance improvement effect of timbre clustering and rhythm pattern matching in ethnic music style recognition, statistical analysis was conducted on the average F1, macro average precision, and macro average recall of four feature modeling strategies to explore the influence of different feature fusion methods on classification stability and discriminative ability. Table 3 showed that the MFCC+SVM model achieved 77.3%, 78.1%, and 76.5% in the three indicators, respectively, and its performance was limited by the discriminative dimension of a single timbre feature. The results of DTW+SVM increased to 82.4%, 83.2%, and 81.6%, respectively, and the introduction of rhythm time structure enabled more accurate characterization of dynamic differences between styles. The PCA+GMM+SVM model

TABLE 2. Quantitative analysis results of separability of timbre, rhythm, and fusion feature space

Feature Type	Inter-Class Distance ↑	Intra-Class Variance ↓	DB Index ↓	Silhouette Coefficient ↑
Timbre Feature Space	1.857	3.412	2.276	0.218
Rhythm Feature Space	3.941	2.093	1.198	0.437
Fused Feature Space	7.926	1.024	0.615	0.708

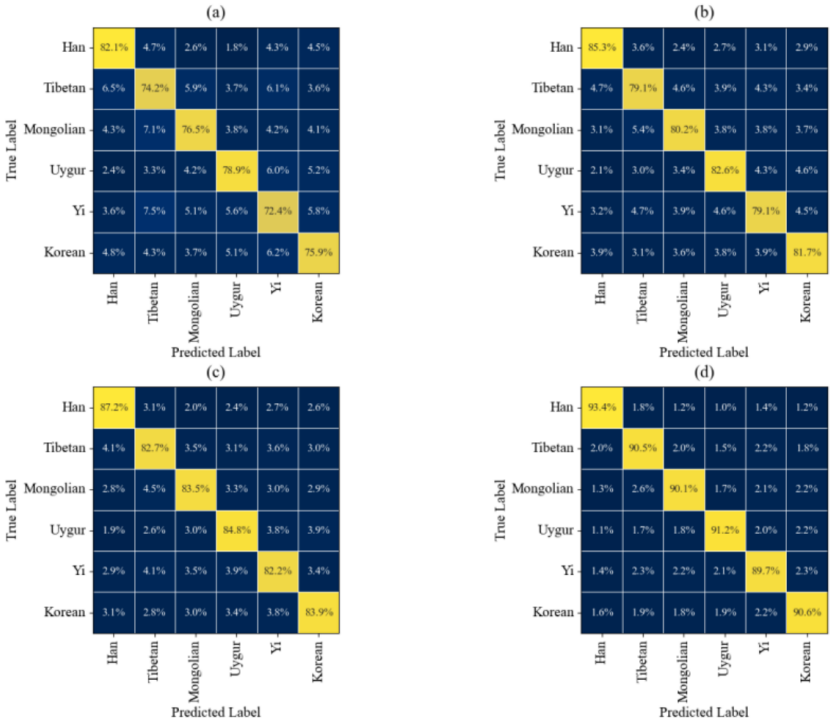

FIGURE 6. Comparison of Confusion Matrix for Ethnic Music Style Classification under Different Feature Modeling Strategies. (a).

further increased these three metrics to 85.2%, 86.1%, and 84.4%, demonstrating that the timbre clustering structure enhanced the internal consistency of multi-instrument signals. The GMM+DTW+SVM fusion model achieved the highest values in all metrics, at 91.3%, 92.0%, and 90.7%, respectively, demonstrating the synergistic advantages of timbre layering and rhythmic features.

METHOD ROBUSTNESS ANALYSIS

In this experiment, this paper investigated the robustness of different feature fusion strategies in the task of folk music style classification, focusing on the number of timbre clusters, rhythmic complexity, and signal-to-noise ratio. Strategy 1, a model that uses only timbre features, relies on MFCC and spectral features after GMM clustering, emphasizing the ability to distinguish instrument timbre. Strategy 2, a model that uses only rhythm features, extracts rhythmic pattern similarity through adaptive beat detection and rhythm template matching, focusing on rhythmic structure. Strategy 3, a timbre and rhythm fusion feature model, combines the probability distribution of timbre clusters with the rhythm matching score to form a multidimensional vector input into a SVM classifier, taking into account both timbre and rhythm information. The robustness analysis results are shown in Figure 7.

The data shows significant strategic differences. As the number of timbre clusters increased, the classification accuracy of strategy three first increased to 78.3% and then slightly decreased to 71.8%.

Strategies one and two decreased from 68.5% to 65.3% and 65.1% to 63.5%, respectively. This indicates that overly fine timbre clustering leads to a decrease in the performance of single feature models, while fused features alleviate the problem of feature dilution. The rhythm complexity increased from 0.1 to 1.0; strategy three decreased from 80.5% to 71.2%; strategy two decreased significantly; strategy one also decreased, but the magnitude was lower than strategy two. This indicated that rhythm features alone were prone to distortion under complex rhythm conditions, while fused features maintained high stability. The signal-to-noise ratio decreased from 30 to -15. As shown in Figure 7, the decrease of strategy three was the smallest from 82.2% to 65.8%. The two strategies had an important decrease under the high noise condition, which demonstrated that one feature was sensitive to the environmental interferences.

TABLE 3. Comparison of Performance Indicators for Ethnic Music Style Classification under Different Feature Modeling Strategies

Feature Modeling Method	Mean F1 (%)	Macro Precision (%)	Macro Recall (%)
MFCC + SVM	77.3	78.1	76.5
DTW + SVM	82.4	83.2	81.6
PCA + GMM + SVM	85.2	86.1	84.4
GMM + DTW + SVM	91.3	92.0	90.7

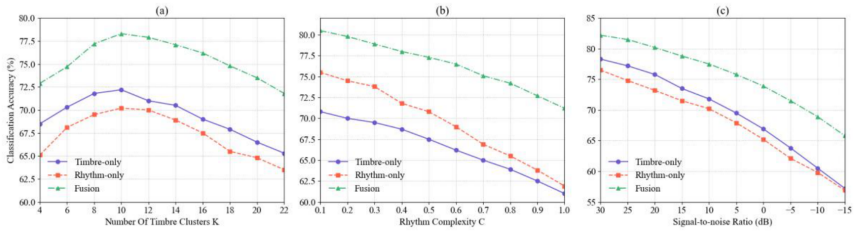

FIGURE 7. Robustness Analysis Line Chart of Ethnic Music Style Classification. (a) Number of timbre clusters, (b) rhythm complexity, (c) signal-to-noise ratio

COMPARISON WITH END-TO-END DEEP LEARNING BASELINES

To examine the performance of the proposed modeling framework within contemporary audio classification paradigms, a variety of end-to-end deep learning models were introduced for comparative analysis while maintaining consistency in data partitioning, feature preprocessing, and training strategies. The comparative models include convolutional neural networks that directly perform two-dimensional convolutional modeling based on logarithmic Mel-frequency spectra, emphasizing hierarchical abstraction of local time-frequency textures; convolutional recurrent neural networks that integrate convolutional feature extraction and temporal recursive modeling, emphasizing implicit modeling of temporal dependencies; recurrent neural networks that use MFCC sequences as input, characterizing overall dynamic changes through temporal state propagation. All these models employ end-to-end training and do not introduce explicit timbre decomposition or rhythm alignment mechanisms, forming a clear contrast in structural assumptions to the hierarchical modeling strategy proposed in this paper. The results are shown in Table 4.

TABLE 4. Performance comparison with end-to-end deep learning baselines

Method	Overall Accuracy (%)	Macro Precision (%)	Macro Recall (%)	F1-score (%)
CNN on log-Mel spectrogram	86.4	87.1	85.6	86.0
CRNN on log-Mel spectrogram	88.2	88.9	87.6	88.1
RNN on MFCC sequence	84.7	85.3	84.1	84.6
Proposed GMM + DTW + SVM	90.9	92.0	90.7	91.3

As shown in Table 4, the convolutional model based on logarithmic Mel spectrum achieved an overall classification accuracy of 86.4% and a macro-average F1 score of 86.0%, demonstrating stable performance in time-frequency local pattern modeling. Introducing a temporal recursive structure further improved the overall accuracy of the convolutional recursive model to 88.2% and the macro-average F1 score to 88.1%, indicating that implicit temporal modeling alleviated the discrimination instability caused by rhythmic variations to some extent. The cyclic model based on MFCC sequences achieved an overall accuracy of 84.7% and a macro-average F1 score of 84.6%, but its overall performance was relatively limited due to the abstraction level of the input representation and its long-range dynamic modeling capabilities. In contrast, the proposed method achieved an overall accuracy of 90.9% and a macro-average F1 score of 91.3%, maintaining higher levels in both macro-average precision and macro-average recall. This demonstrates the enhancing effect of explicit timbre distribution constraints and rhythm alignment mechanisms on classification stability under conditions of multi-source mixing and rhythmic fluctuations, thus verifying the effectiveness of this hierarchical modeling strategy in the current experimental setting.

V. CONCLUSION

This article establishes a collaborative representation framework between timbre structure and rhythm patterns by constructing a classification method for ethnic music styles that combines timbre clustering and rhythm pattern matching. After STFT and MFCC feature extraction of audio, GMM is used to cluster timbres in a reduced dimensional space. The energy envelopes of each timbre cluster are extracted, and the rhythm similarity is calculated using dynamic time warping and ethnic rhythm templates. The timbre probability distribution and rhythm features are fused and input into SVM for classification. The experimental results showed that the clustering purity of GMM in Cluster 2, Cluster 4, and Cluster 6 reached 0.85, 0.88, and 0.89, respectively, with corresponding intra cluster similarity of 0.88, 0.91, and 0.90, reflecting high adaptability to the spectral structure of string and orchestral music. The median rhythm correlation coefficient of the DTW method in the rhythm matching stage was 0.800; the contour coefficient of the fused feature space was increased to 0.708; the inter class distance was increased to 7.926. In the classification experiment, the overall accuracy of the fusion model GMM+DTW+SVM reached 90.9%, with an average F1 score of 91.3%, both significantly higher than the single feature model. The results validated the structural separation ability of timbre

clustering in complex polyphonic signals and the robustness of rhythm pattern matching in time alignment. The fusion of the two effectively alleviated the effects of timbre overlap and rhythm drift. This method has high application value in the fields of digital classification, intelligent retrieval, and cultural inheritance of ethnic music, providing a scalable technical path for multimodal acoustic feature fusion.

AUTHOR CONTRIBUTIONS

Zhen Zhao designed, collected and analyzed the data, and drafted the manuscript. Zhen Zhao conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Zhen Zhao participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by, 2024 Bijie City Social Science Federation Guizhou University of Engineering Science Joint Fund Project,

DOCTORAL RESEARCH START-UP FUND PROJECT

Project Name: "Research on the Integration of National Music Culture and Music Teaching Methods - Taking Bijie City as an Example" Project Number: BSLB-202416

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] L. Liu and H. Liang, "Analysis of the Style Characteristics of Regional Folk Songs and Music Classification Algorithms," Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 29, no. 1, pp. 33-40, 2025, doi: 10.20965/jaciii.2025.p0033.

[2] L. Xie, Y. Wang, and Y. Gao, "Acoustical feature analysis and optimization for aesthetic recognition of Chinese traditional music," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2024, no. 1, pp. 7-20, 2024, doi: 10.1186/s13636-023-00326-2.

[3] X. Wang, L. Wang, and L. Xie, "Comparison and analysis of acoustic features of western and Chinese classical music emotion recognition based on VA model," Applied Sciences, vol. 12, no. 12, pp. 5787-5804, 2022, doi: 10.3390/app12125787.

[4] R. Pandeya Y and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," Multimedia Tools and Applications, vol. 80, no. 2, pp. 2887-2905, 2021, doi: 10.1007/s11042-020-08836-3.

[5] Q. Yue, L. Wang, and J. Luo, "Pattern Recognition Based Music Style Recognition and Teaching Application in Higher Education Music Education," Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, pp. 1-12, 2024, doi: 10.2478/amns-2024-3474.

[6] H. Xue, C. Sun, M. Tang, C. Hu, Z. Yuan, M. Huang, et al., "Effective acoustic parameters for automatic classification of performed and synthesized Guzheng music," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2023, no. 1, pp. 50-62, 2023, doi: 10.1186/s13636-023-00320-8.

[7] M. Chen, "Application of Deep Learning in Music Genre Classification," Frontiers in Artificial Intelligence Research, vol. 2, no. 3, pp. 248-260, 2025, doi: 10.71465/fair331.

[8] Y. Zhang and T. Li, "Music genre classification with parallel convolutional neural networks and capuchin search algorithm," Scientific Reports, vol. 15, no. 1, pp. 9580-9597, 2025, doi: 10.1038/s41598-025-90619-7.

[9] K. Zemlianova, A. Bose, and J. Rinzel, "Dynamical mechanisms of how an RNN keeps a beat, uncovered with a low-dimensional reduced model," Scientific reports, vol. 14, no. 1, pp. 26388-26401, 2024, doi: 10.1038/s41598-024-77849-x.

[10] W. Large E, I. Roman, C. Kim J, J. Cannon, J. Pazdera, L. Trainor, et al., "Dynamic models for musical rhythm perception and coordination," Frontiers in Computational Neuroscience, vol. 17, no. 1, pp. 1151895-1151907, 2023, doi: 10.3389/fncom.2023.1151895.

[11] K. Gourisaria M, R. Agrawal, M. Sahni, and P. Singh, "Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques," Discover Internet of Things, vol. 4, no. 1, pp. 1-24, 2024, doi: 10.1007/s43926-023-00049-y.

[12] E. Setiorini, M. Widjaja, and A. Wicaksana, "Reduced Convolutional Recurrent Neural Network Using MFCC for Music Genre Classification on the GTZAN Dataset," Informatica, vol. 49, no. 17, pp. 1-10, 2025, doi: 10.31449/inf.v49i17.6885.

[13] K. Schulze-Forster, G. Richard, L. Kelley, C. Doire, and R. Badeau, "Unsupervised music source separation using differentiable parametric source models," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, no. 1, pp. 1276-1289, 2023, doi: 10.1109/TASLP.2023.3252272.

[14] G. Zhu, J. Darefsky, F. Jiang, A. Selitskiy, and Z. Duan, "Music source separation with generative flow," IEEE Signal Processing Letters, vol. 29, no. 1, pp. 2288-2292, 2022, doi: 10.1109/LSP.2022.3219355.

[15] C. Cavicchia, M. Vichi, and G. Zaccaria, "Parsimonious ultrametric Gaussian mixture models," Statistics and Computing, vol. 34, no. 3, pp. 108-131, 2024, doi: 10.1007/s11222-024-10405-9.

[16] K. Sreekar and D. Reddy A, "Musical tones classification using machine learning," International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 12, pp. 1004-1007, 2022, doi: 10.22214/ijraset.2022.48084.

[17] Seo Jin Soo, "Brief Paper: Salient Chromagram Extraction Based on the Savitzky-Golay Filter for Cover Song Identification," J Multimed Inf Syst, vol. 9, no. 1, pp. 69-72, 2022, doi: 10.33851/JMIS.2022.9.1.69.

[18] X. Chen, Y. Guo, and J. Na, "Encoder signal-based optimized Savitzky-Golay and adaptive spectrum editing for feature extraction of rolling element bearing under low-speed and variable-speed conditions," ISA transactions, vol. 154, no. 1, pp. 371-388, 2024, doi: 10.1016/j.isatra.2024.07.034.

[19] D. Jiang, "Matching model of dance movements and music rhythm features using human posture estimation," Computational Intelligence and Neuroscience, vol. 2022, no. 1, pp. 7331210-7331222, 2022, doi: 10.1155/2022/7331210.

[20] J. Kraprayoon, A. Pham, and J. Tsai T, "Improving the robustness of DTW to global time warping conditions in audio synchronization," Applied Sciences, vol. 14, no. 4, pp. 1459-1476, 2024, doi: 10.3390/app14041459.

[21] J. Stübinger and D. Walter, "Using multi-dimensional dynamic time warping to identify time-varying lead-lag relationships," Sensors, vol. 22, no. 18, pp. 6884-6899, 2022, doi: 10.3390/s22186884.

[22] Y. Xia and F. Xu, "Study on music emotion recognition based on the machine learning model clustering algorithm," Mathematical Problems in Engineering, vol. 2022, no. 1, pp. 9256586-9256598, 2022, doi: 10.1155/2022/9256586.

[23] S. Hızlısoy, S. Arslan R, and E. Çolakoğlu, "Music Genre Recognition Based on Hybrid Feature Vector with Machine Learning Methods," Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, vol. 38, no. 3, pp. 739-750, 2023, doi: 10.21605/cukurovaumfd.1377737.