A Study on the Classification and Inheritance Genealogy of Ethnic Music Intangible Cultural Heritage Based on Audio Signal Processing

Jing Chen
School of Music, Shandong Women's University, Jinan 250300, Shandong, China
Corresponding author: Jing Chen (e-mail: chezongxing@163.com)

ABSTRACT Accurate style identification, lineage reconstruction, and digital preservation remain critical challenges in safeguarding ethnic music intangible cultural heritage. As audio signal processing and intelligent sensing technologies continue to evolve, efficient feature extraction and pattern recognition methods have become increasingly important for multimedia information analysis and electromagnetic signal processing applications. This study proposes a framework for ethnic music style classification and inheritance genealogy reconstruction based on audio signal processing. Mel-Frequency Cepstral Coefficients (MFCCs) are first extracted to characterize acoustic features, after which a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture is employed for style classification. Subsequently, Dynamic Time Warping (DTW) is adopted to measure the similarity between musical segments, and hierarchical clustering is utilized to reconstruct the lineage network. A comprehensive evaluation dataset containing 2,847 audio samples from 12 ethnic categories is established for validation. Experimental results demonstrate that the proposed approach achieves a style classification accuracy of 94.3% and successfully identifies 89 inheritance branches, accurately reconstructing the evolutionary relationships of ethnic music traditions, including Tibetan and Mongolian genres. The proposed framework provides an effective technical solution for the digital protection and inheritance of ethnic music intangible cultural heritage while offering methodological reference for intelligent acoustic sensing, electromagnetic signal analysis, and multimodal information processing systems.

INDEX TERMS audio signal processing, ethnic music classification, deep learning, electromagnetic signal analysis, MFCC

I. INTRODUCTION

ETHNIC music, an intangible cultural heritage, bears the historical memory and cultural genes of different ethnic groups, yet the preservation and passing on of the cultural heritage are severely challenged. Many valued audio resources are in a disorganized form because of the absence of efficient digital management techniques and style classification is based on manual recognition, and is not only inefficient, but very subjective. The elderly nature of successors has rendered the oral transmission scheme unsustainable and the succession between teachers and students and the development of schools are slowly becoming indistinct [1,2]. These many audio samples are not systematically annotated, limiting the deep use of intangible cultural heritage resources. The question of how to apply technical means to get automated classification and reconstruction of the inheritance genealogy of ethnic music has become an urgent need in the sphere of cultural protection. The machine learning and audio signal processing technologies have offered novel methods of analyzing ethnic music. MFCC characteristics are able to describe timbre and pitch shifts effectively, CNN is capable of identifying local patterns, and LSTM can learn sequence dependencies. Complex musical features recognition ability can be improved with the combination of the two. The DTW algorithm addresses the issue of similarity measurement at varying playing speeds by time alignment and hierarchical clustering can be used to extract hierarchical relationships out of similarity data. The literature has been primarily in terms of popular music genres or individual ethnic music studies, but not a full-scale study of multi-ethnic music and the technical aspects of lineage construction is quite feeble [3,4]. It is a work that builds an audio signal processing-based framework of ethnic music style classification and lineage analysis and has three modules: MFCC features extraction, CNN-LSTM hybrid classification model, and

DTW hierarchical clustering lineage construction. A data set of 12 ethnic groups and 2847 samples was determined and the validity of the approach was tested by use of comparative experiments.

II. RELATED WORK

The method of classifying music styles has been developed into deep learning as opposed to machine learning in the past. Initial research primarily derived time-frequency domain audio characteristics, such as short-time fourier transform coefficients, chromaticity features, rhythmic templates, and applied classifiers, such as SVM and decision trees, to accomplish style recognition. These are based on manually constructed characteristics and do not express complicated musical patterns as well. End-to-end learning has been made possible thanks to the introduction of deep learning technology. Convolutional neural networks automatically learn local features on spectrograms by performing multi-layer convolution operations and recurrent neural networks and variants have the capability to learn the temporal structure of musical sequences. This has been augmented by the use of attention mechanisms to help the model pay attention to important musical passages [5,6]. Ethnic music studies have primarily been done in single category timbre or instrument recognition. The overall classification system of the ethnic groups is not yet complete and the techniques of extracting the features are not yet adequate in modelling ethnic music-specific features like different tones and decorations. In the music information retrieval, similarity calculation of music and the relationship analysis with inheritance is also a significant issue. Euclidean distance-based similarity measures assume that the sequences are the same length, and are strictly aligned, which is hard to deal with tempo variation and rhythmic freedom in real performances. Through dynamic programming, the DTW algorithm determines the optimum time alignment path, and is therefore successful at solving the matching problem in variable-tempo performances, but it is computationally expensive. Empirical research on music inheritance connections usually uses a social network analysis or phylogeny, building inheritance maps using metadata, but does not directly analyse audio material. Clustering-based methods have been used in the discovery of music genres, although little has been done to explore mining the hierarchical nature of inheritance lineages [7,8]. The current processes are not effective at integrating style classification and lineage analysis and have not been able to deliver the dual requirements of intangible cultural heritage protection of accurate identification and tracing of relationships.

III. METHODS

A. DATA ACQUISITION AND PREPROCESSING

In this study, a dataset of 2,847 audio samples of 12 ethnic groups was created, including representative types of ethnic music, like Tibetan, Mongolian, Uyghur, Miao, and Dong. The audio resources were found on the National Intangible Cultural Heritage Digital Platform, provincial cultural center archives, and field survey records. Sampling rate was constant 44.1 kHz, bit depth 16 bits and format mono.

There are three major steps of data preprocessing. Spectral subtraction and Wiener filtering are employed to reduce noise in the environment during field recording, and a noise signal threshold of over 20dB is established. Given the difference between the gain of various recording devices, peak normalization is applied to bring all audio amplitudes to -3dB to counteract any discrepancy in the dynamic range that would have impacted feature extraction. In audio segments with mixed vocals and instrumental music, the independent component analysis is performed to isolate the vocals and the accompaniment signal and retain the pure melody signal to be further analyzed [9,10].

To capture the transition aspects of music segments, audio slicing uses a sliding window whose window length is 3 seconds and a step size is 1.5 seconds. They are pad padded to the standard length in short segments less than 3 seconds in length. Table 1 details the distribution of the dataset.

Table 1 shows that the number of music samples from different ethnic groups is relatively evenly distributed, with the sample size for each ethnic group ranging from 187 to 318 pieces. Although the absolute sample size per group is finite, a stratified sampling strategy was employed to ensure that key regional sub-genres (e.g., Kham vs. Amdo in Tibetan music) were represented. To mitigate the risk of overfitting to specific recordings, data augmentation techniques and cross-validation were applied, ensuring the model identifies broad stylistic markers rather than recording-specific noise. Musical expression of different ethnic groups differs with the Miao, Dong, and Li ethnic groups dominated by singing styles (over 68) and the Uyghur and Kazakh ethnic groups dominated by instrumental styles (more than 50). Regarding the mean length, the audio pieces of the Mongolian, Hui and Kazakh ethnic groups are longer, all more than 190 seconds, which is connected with the narration nature of their musical frameworks. The sources of data include a variety of channels, such as official cultural institutions, academic research units, and field surveys, which guarantees the authority and diversity of the samples.

Following the preprocessing described in Section 3.1, all subsequent feature extraction and model training were conducted using the pure melody signals isolated through ICA. This ensures that the 45-dimensional feature representation accurately captures the melodic and timbre characteristics of the ethnic music, independent of background noise or accompaniment variations.

The MFCC feature extraction algorithm entails the framing of the processed audio signal with a Hamming window frame length equal to 25 milliseconds and frame shift equal to 10 milliseconds. The time-domain signal is then transformed to the frequency domain in a Fast Fourier Transform (FFT). A 40-mel filter bank filters the frequency-domain signal, but the filter center frequencies are spread non-linearly based on the Mel scale. The resolution in the

TABLE 1. Statistical distribution of samples in the ethnic music dataset

Ethnic categories	Sample size	Average duration (seconds)	Percentage of performance styles (%)	Instrument form percentage (%)	Data source
Tibetan	318	187.3	62.3	37.7	Tibet Cultural Center/Field Recording
Mongolian	276	203.5	55.1	44.9	Inner Mongolia Art Research Institute
Uighur	294	176.8	48.6	51.4	Xinjiang Intangible Cultural Heritage Center
Miao ethnic group	241	154.2	71.4	28.6	Guizhou Nationalities Museum
Dong ethnic group	228	168.9	68.9	31.1	Guangxi Cultural Archives
Yi ethnic group	253	181.6	59.7	40.3	Yunnan Arts University
Zhuang ethnic group	219	159.4	64.8	35.2	Guangxi Intangible Cultural Heritage Protection Center
Hui people	187	192.7	52.4	47.6	Ningxia Cultural Center
Tujia people	205	163.5	66.3	33.7	Hunan Institute of Nationalities
Kazakh	234	198.2	46.2	53.8	Xinjiang Musicians Association
Dai ethnic group	196	171.8	58.7	41.3	Yunnan Nationalities Cultural Palace
Li ethnic group	196	147.6	72.4	27.6	Hainan Intangible Cultural Heritage Center

low-frequency area is greater than the resolution in the high-frequency area, which is consistent with pitch perception by the human ear. A Discrete Cosine Transform (DCT) is then applied to the logarithmic energy spectrum to take the first 13 dimensions of coefficients. At the same time, both the first-order and second-order difference coefficients (both 13-dimensional) are computed, creating a 39-dimensional MFCC feature vector. In order to further improve the timbre variation properties of the ethnic music, six more aspects are elicited such as zero-crossing rate, spectral centroid, and spectral roll-off point which finally produces 45-dimensional feature representation.

The CNN-LSTM hybrid network uses a layered structure. The convolutional layers are made up of three sets of convolutional units of 3×3, 5×5 and 7×7 kernel sizes consisting of 64 filters each to represent local patterns at varying time scales. The two 2×2 windows of max pooling layers are used to reduce the feature dimensionality, and the batch normalization layers are used to speed up the convergence of the model. The convoluted feature maps of the convolutional layers are fed into a bidirectional LSTM network which has two layers with each having 128 hidden units. It uses forward and backward propagation to learn long-term dependencies in music sequences. The LSTM layer outputs are transformed to a 12-dimensional classification space through fully connected layers and the outputs of the softmax activation function are the probability distribution of each ethnic category. The model is based on a cross-entropy loss function, and the weights are optimized by Adam. Learning rate will be 0.0001, and the training epochs will be 50. To avoid overfitting, the dropout ratio will be 0.5. This classification module establishes the categorical boundaries necessary for further analysis. Once the styles are accurately identified, the framework transitions from global classification to granular structural comparison, using the high-dimensional features extracted by the CNN-LSTM model as the input for the subsequent lineage mapping process.

B. A GENEALOGY CONSTRUCTION METHOD BASED ON DTW AND HIERARCHICAL CLUSTERING

The DTW algorithm compares the segments of music by building a cumulative distance matrix to determine their similarity. The logical premise of using similarity for lineage reconstruction is based on the ethnomusicological theory that core melodic and rhythmic structures act as “acoustic fingerprints” which persist through oral transmission. By quantifying these structural invariants, the cumulative distance matrix provides an objective, data-driven foundation for tracing evolutionary paths that complements traditional historical narratives. Because signal lengths of the feature sequences of two audio segments can be different, DTW applies dynamic programming to identify the optimal time alignment path. Euclidean distance is the metric of distance used in calculating the difference between feature vectors. The cumulative distance table is filled gradually starting at the starting point and moving towards the final point and the value at each point is a sum of the current frame distance and a minimum of the cumulative distance of the three neighboring positions. Path constraints are used to limit the matching direction to avoid ill-conditioned alignment, and the slope constraint is established in the range 0.5 to 2. The cumulative distance value at the final position is normalized and it serves as the measure of similarity; the lower the value the more similar the music segments. All pairs of samples are computed by the DTW distance to form a 2847×2847 similarity matrix that is symmetric. The hierarchical clustering is an agglomerative technique that builds the musical traditions lineage. First, independent nodes are considered to be each audio track. A similarity matrix is used to calculate the distance between nodes, and the Ward connection criterion to merge the two nearest nodes. This criterion reduces the amount of squares in the cluster and it guarantees compactness intra-clusters post-merging. The merging process is a recursive process, where the distance matrix is updated at each iteration and the merging order and distance value is noted. A dendrogram is used to show the hierarchical form with the horizontal axis displaying the sample number and the vertical axis displaying the distance threshold during merging. Multi-level clustering is achieved by varying the truncation thresholds. To determine the optimal threshold, a sensitivity analysis was performed using the Silhouette Coefficient across a range of 0.1 to 0.6. The threshold of 0.35 was selected as it yielded the highest mean silhouette score (0.72) and showed the strongest alignment with historical lineage records validated by ethnic musicology experts. At this threshold, 89 lineages are found, each of which is a cluster of works sharing musical features. To avoid circular logic, the clustering process was performed strictly on extracted acoustic features, with the model blinded to any geographical or ethnic metadata. Post-clustering, the analysis of metadata regarding the geographical distribution of samples in each lineage, data on inheritors, and techniques of singing was conducted solely for external validation to confirm that the identified clusters show the evolutionary route of ethnic music. The Mongolian long song displays a distinct eastern and western division of Inner Mongolia, whereas Tibetan string dance music has created local variations of the music such as Changdu and Kangding.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL DATASET

The experiment used a part of the ethnic music of the FMA (Free Music Archive) dataset and a homemade dataset of Chinese ethnic music. The FMA data has 106,574 music tracks, including the audio samples of various ethnic groups in the world music category, and its sampling rate is 44.1 kHz. It gives uniform metadata annotations which can be compared and validated across studies. The self-constructed dataset has 2,847 journeys of Chinese ethnic music that deals with the lack of such journeys in the FMA dataset. The experimental setup was set up on an NVIDIA Tesla V100 that had 32GB of VRAM and CUDA 11.2 acceleration framework. The data were separated into training, validation and test sets of 7:2:1 ratio and 5-fold cross-validation was performed to test the generalization capability of the model.

B. STYLE CLASSIFICATION PERFORMANCE ANALYSIS

Five baseline methods were set up, including traditional machine learning methods SVM and Random Forest, and deep learning methods CNN alone, LSTM alone, and a simple fully connected neural network. All methods used the same MFCC feature input, and the training parameters were kept consistent to ensure fair comparison. Table 2 shows the classification performance metrics of each method on the test set.

Table 2 shows that the CNN-LSTM hybrid model outperforms other methods on all evaluation metrics. Compared to the SVM method, the F1 score improved from 0.74 to 0.94, demonstrating that the deep learning architecture significantly surpasses traditional machine learning in modeling the complex features of ethnic music. The accuracy of the standalone CNN model was 87.5%, the standalone LSTM reached 89.2%, while the hybrid model further improved to 94.3%, indicating that the local features extracted by the convolutional layers and the sequence dependencies captured by the LSTM are complementary. Random forest improved performance by 3.3 percentage points compared to SVM, but still lagged behind the hybrid model by 15.4 percentage points. In terms of inference time, SVM was the fastest but had the lowest accuracy, while the hybrid model maintained a reasonable inference speed of 0.23 seconds while maintaining high accuracy.

C. RESULTS OF GENEALOGY RECONSTRUCTION

The clustering results were verified using manually labeled inheritance relationship data, with samples from the same inheritor, the same regional school, or the same musical form used as the true labels. Table 3 shows

the identification results and regional distribution characteristics of the inheritance branches of major ethnic groups.

Table 3 shows that the 89 lineages identified by the system exhibit significant regional clustering characteristics. The 11 Tibetan lineages correspond to the musical differences of the three major dialect regions of Kham, Amdo, and Ü-Tsang, with the Kham stringed instrument lineage having a sample size of 47 pieces, reflecting the high activity level of this type of musical transmission. The 9 Mongolian lineages successfully distinguished the stylistic differences between eastern and western Inner Mongolia, with Hulunbuir long song and Alashan long song belonging to different lineages, verifying the influence of geographical isolation on musical evolution. Among the 8 Uyghur lineages, Kashgar Muqam and Turpan Muqam show significant differentiation in melodic features. Miao flying songs have formed two independent lineages in southeastern Guizhou and western

TABLE 2. Comparison of the performance of different methods in classifying ethnic music styles

Method	Accuracy (%)	Accuracy (%)	Recall rate (%)	F1 value	Reasoning time (seconds/entry)
SVM	75.6	73.8	74.2	0.74	0.08
Random Forest	78.9	77.5	78.1	0.78	0.12
Fully connected network	82.3	81.7	81.9	0.82	0.15
CNN	87.5	86.9	87.2	0.87	0.19
LSTM	89.2	88.6	88.9	0.89	0.27
CNN-LSTM hybrid model	94.3	93.8	94.1	0.94	0.23

TABLE 3. Statistics on the Identification Results of Musical Heritage Lineages of Various Ethnic Groups

Ethnic categories	Identification coefficient	Main geographical distribution	Representative musical forms	Maximum branch sample size	Minimum branch sample size
Tibetan	11	Tibet, Qinghai, Sichuan, Gansu	String, Dui Xie, Nang Ma	47	12
Mongolian	9	Eastern and western Inner Mongolia, Mongolia	Long song, throat singing, and morin khuur music	52	15
Uighur	8	Southern Xinjiang, Northern Xinjiang, Eastern Xinjiang	Muqam, Meshrep	58	18
Miao ethnic group	7	Guizhou, Hunan, Guangxi	Flying Songs, Wandering Songs, Drinking Songs	49	14
Dong ethnic group	6	Guizhou, Guangxi, Hunan	Grand Song, Pipa Song, Niutuiqin Song	54	19
Yi ethnic group	8	Yunnan, Sichuan, Guizhou	Hai Cai Qiang, A Xi Tiao Yue	44	16
Zhuang ethnic group	6	Guangxi and Yunnan	Mountain songs, row songs	51	17
Hui people	5	Ningxia, Gansu, Qinghai	Flowers, Banquet Song	56	20
Tujia people	7	Hunan, Hubei, Chongqing	Waving Hand Song, Weeping Wedding Song	42	13
Kazakh	7	Northern and central Xinjiang	Dombra singing and playing, Aken singing and playing	48	16
Dai ethnic group	6	Xishuangbanna and Dehong in Yunnan	Zanha Tune, Peacock Dance	45	18
Li ethnic group	5	Cities and counties in Hainan	March 3rd Love Song, Labor Song	53	21

Hunan, with pitch variation patterns exhibiting regional characteristics.

V. CONCLUSION

The framework to classify and analyze lineage of ethnic music styles according to audio signal processing effectively solve the technical problems of digital preservation of intangible cultural heritage. MFCC features are used to input CNN-LSTM hybrid model to realize accurate identification of musical styles from 12 ethnic groups. DTW algorithm and hierarchical cluster method are used to reconstruction multi-level lineage network successfully, and the regional differentiation and development of different style are revealed. The technical system based on lineage reconstruction of ethnic music is applied to the archive management of intangible cultural heritage, cultivation of inheritor and digital display of cultural heritage. Currently, the method still has the following problems: the classification accuracy of musical style from small sample ethnic group needs to be improved; the ability to identify cross-ethnic musical fusion is weak; the impact of singing style of inheritor should be quantified. In the future, transfer learning technology will be used to extend to more ethnic categories; multi-mode information including lyrics and performance technique will be integrated; the knowledge graph of inheritance will be established and updated dynamically; and the visualization analysis platform of cultural worker will be developed.

AUTHOR CONTRIBUTIONS

Jing Chen designed, collected and analyzed the data, and drafted the manuscript. Jing Chen conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jing Chen 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 received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] W. Wang and M. Sohail, "Research on Music Style Classification Based on Deep Learning," Computational and mathematical methods in medicine, vol. 2022, 3699885, 2022, doi: 10.1155/2022/3699885.
[2] Q. Ning and J. Shi, "Artificial Neural Network for Folk Music Style Classification," Mobile Information Systems, 2022, doi: 10.1155/2022/9203420.
[3] R. Tang, M. Qi, and N. Wang, "Music style classification by jointly using CNN and Transformer," Proceedings of the 2024 16th International Conference on Machine Learning and Computing, pp. 707-712, 2024, doi: 10.1145/3651671.3651696.
[4] K. Zhang, "Music Style Classification Algorithm Based on Music Feature Extraction and Deep Neural Network," Wirel. Commun. Mob. Comput, vol. 2021, pp. 9298654: 1-9298654: 7, 2021, doi: 10.1155/2021/9298654.
[5] I. Medicine C A M M, "Retracted: Research on Music Style Classification Based on Deep Learning," Computational and mathematical methods in medicine, vol. 2022, 9823985, 2022, doi: 10.1155/2022/9823985.
[6] L. Xu, "Research on Music Style Classification and health care Based on Neural Network," Journal of Commercial Biotechnology, 2022, doi: 10.5912/jcb1235.
[7] C. Wang and X. Geng, "Simulation of music style classification model based on support vector machine algorithm," Proceedings of SPIE, vol. 12703, pp. 8, 2023, doi: 10.1117/12.2682982.
[8] Q. Ma, C. Yuan, W. Zhou, et al., "Beyond Statistical Relations: Integrating Knowledge Relations into Style Correlations for Multi-Label Music Style Classification," ACM, 2020, doi: 10.1145/3336191.3371838.
[9] X. Zhao and Q. Tuo, "Classification method of popular music score style based on SVM," International Journal of Reasoning-based Intelligent Systems, 2023, doi: 10.1504/ijris.2023.10052718.
[10] C. Liu, C. Wang, and Z. Hu, "Intelligent Classification Model of Oroqen Music Style Based on Deep Learning," 2024 International Conference on Artificial Intelligence, Deep Learning and Neural Networks (AIDLNN), pp. 147-151, 2024, doi: 10.1109/aidlnn65358.2024.00031.