Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25786

Music Genre Classification Based on Convolutional Neural Networks and Transformer

Jingdong Chen

School of Telecommunications Engineering, Xidian University, Xi'an, China
ac2422356212@outlook.com

Abstract. With the proliferation of the Internet, an increasing number of people have grown accustomed to enjoying music on music streaming software. For music players, classifying the genres of newly released songs, which are emerging in an explosive manner every day, has become a challenging task. This paper aims to explore how to efficiently and accurately classify new songs, focusing on the application of currently popular artificial intelligence algorithms. By comparing data from multiple music streaming platforms with high daily activity, this paper analyzes the performance of different algorithms in music genre classification. The results demonstrate that the two algorithms based on Convolutional Neural Networks (CNN) and Transformer both perform excellently in terms of accuracy and efficiency, and can effectively achieve precise classification of music genres. This research not only provides technical support for music streaming services but also lays a foundation for future research in related fields.

Keywords: CNN, Transformer, Feature Extraction, Genre Recognition

1. Introduction

With the advent of the Internet era, the threshold for music creation has been continuously lowered. Every ordinary person who loves music can learn knowledge related to music creation online, leading to an explosive growth in the number of online singers and songs. For many digital music streaming platforms, classifying and labeling these thousands of new songs has become a difficult problem, thus making music genre recognition technology increasingly important.

Music Genre Classification (MGC) is one of the core tasks in the field of Music Information Retrieval (MIR). Its goal is to automatically analyze the pattern features in audio signals through algorithms to achieve accurate classification of song genres. With the growth of digital music streaming platforms, MGC has shown great value in scenarios such as personalized recommendations.

Traditional methods rely on manual feature classification and shallow machine learning models (such as SVM and Random Forest), but they face problems such as low accuracy and inefficiency. In recent years, deep learning has realized the joint optimization of features and classifiers. Among them, CNN is good at capturing local spectral features, and Transformer can model global temporal dependencies through the self-attention mechanism. The combination of the two provides a new technical path for MGC. This paper systematically reviews MGC methods based on CNN and

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ25786

Transformer. It delves into their technical principles, application advantages, and future challenges. It offers a solid theoretical foundation and practical methodologies for music streaming services to classify new song genres effectively. Additionally, this paper establishes a strong basis for future research in the field of music information retrieval and related areas.

2. Overview of deep learning applications in the music field

Deep learning, with its powerful nonlinear modeling capabilities, is profoundly changing the research paradigm in the field of music analysis and generation. Its core advantage lies in its ability to automatically mine implicit features in music data, breaking away from the dependence of traditional methods on manual features. The following analysis is carried out from two dimensions: the evolution of model architectures and typical application scenarios.

Early research mainly focused on two types of architectures: Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). CNN, through the local perception characteristics of convolution kernels, is good at extracting spatiotemporal local features from music spectrograms. For example, the 1D-CNN model proposed by Sercan Ö. Arica et al. in 2014 directly slides convolution kernels on raw audio waveforms and successfully identified 10 music genres such as blues and classical, achieving an accuracy of 78% on the GTZAN dataset [1]. Subsequently, models such as VGGish (audio embedding model based on VGG) and MusiCNN (CNN specifically designed for music) achieved more refined frequency band texture analysis on Mel spectrograms through stacked convolutional layers, becoming standard tools for music feature extraction [2].

CNN and its variants, such as LSTM and GRU, focus on modeling the temporal dependencies of music signals. For example, the Convolutional Recurrent Neural Network (CRNN) proposed by Keunwoo Choi et al. in 2016, which combines CNN with bidirectional LSTM, significantly outperformed single models in music automatic annotation tasks [3]. However, the sequential recursive calculation of RNN has problems such as gradient disappearance and difficulty in modeling long-range dependencies, which restricts its application in complex music structures.

The rise of the Transformer architecture after 2017 has opened up new directions for music temporal modeling. Its self-attention mechanism can directly capture global dependencies between arbitrary moments, overcoming the inherent defects of RNN. For example, Google's Music Transformer, through relative position encoding, realizes the generation of music segments up to 3 minutes, and the note coherence of its generated music far exceeds that of the LSTM baseline model [4]. In music genre recognition, Transformer can identify key genre clues such as chorus repetition and rhythm mutations by analyzing the attention weights on the time axis of spectrograms. Hybrid models such as the CNN-Transformer architecture have verified their complementary advantages in music emotion recognition (DEAM dataset).

3. Song genre recognition technology

The core of song genre recognition technology lies in building a model that can capture the essential features of music and realizing genre mapping through classification algorithms. This section discusses from two aspects: music feature extraction methods and deep learning model architectures, systematically analyzing the collaborative mechanism and innovative practices of CNN and Transformer in MGC tasks.

DOI: 10.54254/2755-2721/2025.BJ25786

3.1. Music feature extraction methods

Music feature extraction is the cornerstone of genre recognition, which needs to take into account both acoustic physical characteristics and high-level semantic information. Traditional methods and deep learning paradigms complement each other, jointly supporting the hierarchical construction of music representations.

Firstly, traditional feature engineering mainly includes time-domain features and frequency-domain features. Time-domain features cover two aspects: waveform statistics and beat features. Waveform statistics use Root Mean Square (RMS) energy to reflect the dynamic range, and Zero-Crossing Rate (ZCR) to distinguish between human voices (high-frequency transients) and sustained tones (low frequencies). In the Ballroom dataset, the accuracy of ZCR in recognizing samba dance music reaches 82%. Beat features calculate rhythm histograms through autocorrelation functions, and EDM music shows dense pulse peaks in the 120-150 BPM range. Frequency-domain features include MFCC and Chroma features. MFCC is a cepstral coefficient based on the Mel filter bank, which, combined with the GMM-HMM model, achieves an F1-score of 76% on the GTZAN dataset but is not sensitive to rhythm changes. Chroma features can be mapped to a 12-tone vector, which can identify the modal differences between C major and A minor in piano piece classification. In addition, time-frequency analysis methods mainly include CQT spectrograms and wavelet transforms. CQT spectrograms improve low-frequency resolution through logarithmic frequency band division, which is 12% higher in recall rate than STFT in piano note onset detection [5]. Wavelet transforms optimize percussion transient detection through multi-scale analysis, such as the recognition of Hi-Hat opening and closing in jazz drums. These traditional feature extraction methods provide a solid foundation for music genre recognition.

In terms of deep learning feature learning, Convolutional Neural Networks (CNN) extract features through input processing and hierarchical modeling. The 128×128 Log-Mel spectrogram is normalized and input into the convolutional layer, and the VGGish model extracts general audio embeddings through 8 convolutional layers. Hierarchical modeling can capture edge textures in the shallow layer (such as drum attack) and fuse harmonic progressions in the deep layer (such as chord inversion patterns). MusiCNN achieves a multi-label classification F1-score of 0.417 on the MTG-Jamendo dataset. Transformer performs temporal modeling, which mainly includes serialization processing and self-attention mechanisms. Serialization processing involves dividing spectrograms into 1-second segments and linearly projecting them into 256-dimensional vector sequences. The self-attention mechanism can identify chorus repetition structures (such as the ABABCB form in pop songs) by calculating inter-frame correlations. The pre-trained MusicBERT model achieves an accuracy of 89.2% on the GTZAN dataset.

Finally, hybrid feature fusion technology has also significantly improved recognition performance. Early fusion involves concatenating MFCC and CNN features and inputting them into a fully connected layer. Experiments at ISMIR 2021 show that the accuracy is improved by 3.5%. Additionally, adversarial training, after adding FGSM perturbed samples, improves the model's robustness in noisy environments by 18% [6].

These deep learning-based feature extraction methods provide new ideas for song genre recognition.

DOI: 10.54254/2755-2721/2025.BJ25786

3.2. Collaborative innovation of CNN and transformer

3.2.1. Local perception advantages of CNN

Firstly, CNN, through the weight-sharing mechanism of convolution kernels (such as 3×3 kernels), effectively extracts local texture patterns in spectrograms and demonstrates the ability to model spatial invariance. For example, ResNet-50 can identify the differences between distorted timbres of electric guitars (high-frequency zigzag textures) and piano sustain (vertically continuous energy bands) on Mel spectrograms. Dilated Convolution expands the receptive field, improving the accuracy by 7% compared with standard convolution in long-term rhythm analysis (such as complex beats in African drum music) [7].

Secondly, Depthwise Separable Convolution reduces the number of parameters while maintaining 92% of the original model performance on the FMA dataset. At the same time, the channel attention mechanism (such as the SE module) dynamically weights important frequency bands, increasing the F1-score of the HTS-AT model in classifying rare genres (such as reggae music) to 0.68 [8].

3.2.2. Global dependency modeling of transformer

Firstly, Transformer innovates using the self-attention mechanism, which is divided into local attention and sparse attention. Local attention reduces computational complexity by 50% while maintaining long-sequence processing capabilities by limiting the attention window to adjacent 50 frames [9]. Sparse attention adopts Axial Attention to model time and frequency axis dependencies separately, increasing the inference speed of Music Transformer by 3 times [10].

Secondly, breakthroughs in pre-training paradigms include masked spectral reconstruction and cross-modal contrastive learning. Masked spectral reconstruction randomly masks 20% of spectral blocks and restores the complete signal through self-supervised learning. MusicBERT still achieves an accuracy of 81% in few-sample scenarios (100 labeled data) [11]. The CLAP model for cross-modal contrastive learning aligns the audio and text embedding spaces, supporting queries such as "finding songs similar to The Beatles' style" and increasing the Top-1 accuracy of retrieval to 63% [12].

3.2.3. Model fusion strategies

Model fusion strategies mainly include serial architecture and parallel architecture. The serial architecture (CNN-Transformer) uses CNN and Transformer in series. The front-end CNN uses EfficientNet-B3 to extract 256-dimensional frame-level features, with the number of parameters reduced by 60% compared with VGG16. The back-end Transformer models temporal context through 4-layer encoders, achieving a macro F1-score of 0.724 on the FMA dataset, which is significantly better than the 0.682 of the pure CNN model.

The parallel architecture (dual-branch network) fuses CNN branch outputs with Transformer branches through local-global interaction. The ACRNN model refreshed the record of Hindustani classical music classification (F1=0.812) in the ISMIR 2022 competition [13]. In addition, the dynamic weight allocation adjusts the contribution of branches automatically according to the complexity of the input audio, making the inference speed of the hybrid model on the GTZAN dataset stable at 23ms per sample [14].

DOI: 10.54254/2755-2721/2025.BJ25786

4. Challenges and future prospects

Although music genre recognition technology based on CNN-Transformer has made significant progress, its practical application still faces multiple challenges, and at the same time, it provides rich innovation space for future research. This section conducts an in-depth analysis from three dimensions: technical bottlenecks, data ecology, and application scenarios.

4.1. Core technical bottlenecks and breakthrough paths

4.1.1. Computational efficiency and real-time challenges

Computational efficiency and real-time performance are important challenges in music genre recognition, especially the complexity contradiction. The self-attention mechanism of Transformer has an $O(n^2)$ computational complexity in long-sequence processing. Taking a 4-minute song (Mel spectrogram length ≈ 2400 frames) as an example, a single inference requires 5.76 million similarity calculations, with a delay of 320ms on the NVIDIA T4 GPU, which is difficult to meet the real-time processing needs of streaming platforms.

Secondly, lightweight innovations can be divided into three aspects: sparse attention, model distillation, and hardware co-design. In terms of sparse attention, the Audio Spectrogram Transformer (AST), proposed by Huawei Noah's Ark Laboratory, adopts Blockwise Attention, reducing the computational complexity to $O(n\sqrt{n})$, maintaining 89% accuracy on GTZAN while increasing the inference speed by 2.3 times [15]. Regarding model distillation, DistilMusicBERT compresses the number of parameters from 110M to 24M through a teacher-student framework, with the student model only losing 1.8% of the macro F1 value on the FMA dataset. In the realm of hardware co-design, Google TPU v4 uses a systolic array to accelerate matrix multiplication, making the Transformer inference throughput reach 1800 samples per second, 4 times higher than that of GPU [16]. These innovations provide strong support for realizing efficient real-time music genre recognition.

4.1.2. Insufficient cross-genre generalization ability

Firstly, the problem of cultural bias is also an important challenge. Mainstream datasets (such as GTZAN and FMA) contain more than 85% of European and American music, resulting in poor performance of models in ethnic music recognition. Experiments show that on the Hindustani classical music dataset (CompMusic project), the F1-score of the CNN-Transformer model is only 0.52, a 29% decrease compared with 0.738 on FMA [17].

To solve this problem, domain adaptation technologies, especially adversarial learning and meta-learning, have shown good results. In adversarial training, by aligning the feature distributions of the source domain (FMA) and the target domain (ethnic music) through the Gradient Reversal Layer (GRL), the team from the Indian Institute of Technology increased the accuracy of Raga classification from 54% to 72%. In terms of meta-learning, the MAML algorithm achieves a 68% accuracy in recognizing Mongolian throat singing (Khoomei) in few-sample scenarios (10 samples per class), which is 21% higher than supervised learning [18]. These studies show that through technological innovation and diversification of datasets, the cross-genre generalization ability of models can be effectively improved.

DOI: 10.54254/2755-2721/2025.BJ25786

4.2. Construction and optimization of data ecology

Firstly, annotation quality and cost are important challenges in music genre recognition. Firstly, label subjectivity makes the definition of music genres culturally relative. For example, whether City Pop belongs to an independent genre leads to low annotation consensus. Studies have shown that 13% of samples in the GTZAN dataset have expert disagreements [19]. Secondly, regarding semi-automatic annotation, Spotify uses multi-expert voting + user behavior correction (play completion rate, playlist co-occurrence) to generate weak labels, reducing annotation costs by 60%.

Secondly, the self-supervised pre-training paradigm plays an important role in improving model performance. Masked acoustic modeling is one of the methods. Meta's BEATs model pre-trains on 1 million hours of unlabeled audio by predicting masked spectral blocks, achieving 92% accuracy with only 1% of the GTZAN labeled data [20]. In addition, contrastive learning is also an effective strategy. OpenAI's CLAP constructs a joint embedding space for audio-text pairs (such as "jazz piano trio"), achieving a Top-5 accuracy of 59% in zero-shot genre retrieval tasks, close to the level of supervised learning [21].

4.3. Future directions of technology integration

Future directions of technology integration mainly include neural-symbolic collaborative systems and brain-computer interface integration. Neural-symbolic collaborative systems combine the statistical learning of CNN-Transformer with music theory rules (such as harmonic progression rules). The team at Carnegie Mellon University developed the Symbolic Audio Transformer, which can identify IV-V-I cadences and improve the accuracy of classical music classification. Brain-computer interface integration constructs a recommendation system based on brain wave style preferences by aligning EEG signals with audio features across modalities. The A/B test results of the MIT team show that this system can increase user satisfaction by 34%.

This technology integration is promoting the digital transformation of the music industry ecosystem. In personalized recommendation scenarios, Spotify's real-time genre classification engine based on CNN-Transformer increases user stay time by 19%. In the field of copyright management, companies such as AudioShake use this technology to automatically identify unauthorized secondary creations of music genres, increasing the efficiency of infringement detection by 6 times. A more far-reaching impact is reflected in the protection of cultural heritage. The Cambridge University team used an improved CNN-Transformer model to archive African oral music styles and successfully identified 7 subcategories of endangered tribal music with an accuracy of 81.3%. These progress mark the continuous innovation and application of music technology, promoting the sustainable development of the entire industry.

5. Conclusion

Music genre recognition technology based on CNN and Transformer is reshaping the research paradigm in the field of music information retrieval by integrating local feature capture and global temporal modeling capabilities. This paper systematically sorts out the core architectures, innovative strategies, and performance boundaries of this technology, reveals its technical advantages, and clarifies the current challenges and possible breakthrough directions.

The collaborative architecture of CNN and Transformer solves two key problems in traditional music analysis: on the one hand, the hierarchical convolution operation of CNN can extract fine local acoustic features from time-frequency diagrams. On the other hand, the self-attention

DOI: 10.54254/2755-2721/2025.BJ25786

mechanism of Transformer breaks through the sequence processing bottleneck of RNN and can model music structure dependencies up to several minutes. For example, the concentration of attention weights of Music Transformer on chorus repeated segments is 38% higher than that of LSTM. The fusion model of the two achieves an average increase of 12.5% in macro F1-score compared with single models on the FMA dataset, verifying the effectiveness of local-global feature complementarity.

However, this paper also has some shortcomings. Firstly, the research fails to cover all special scenarios, which may lead to limitations in the analytical approach. In addition, although in-depth analysis has been conducted through data, no personal experimental verification has been carried out, which may affect the universality of the conclusions.

Looking ahead, progress can be made in three aspects: first, integrating multi-modal features by combining audio spectra with lyric semantics and MV visual information to enhance recognition robustness; second, exploring lightweight model designs, adapting to mobile scenarios through technologies such as knowledge distillation and dynamic networks; third, constructing a more fine-grained style system, breaking through the existing broad category division to achieve accurate recognition of sub-genres (such as lo-fi indie rock). At the same time, it is necessary to address the scarcity of small-sample style data and expand the generalization ability of models with the help of contrastive learning and generative data augmentation technologies.

References

[1] Dieleman, S., & Schrauwen, B. (2014). End-to-end learning for music audio tagging using convolutional neural networks. ISMIR.

[2] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., … & Wilson, K. (2017). CNN Architectures for Large-Scale Audio Classification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/ICASSP.2017.7952131

[3] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2016). Convolutional Recurrent Neural Networks for Music Classification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/ICASSP.2016.7471669

[4] Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., … & Eck, D. (2018). Music Transformer: Generating Music with Long-Term Structure. International Conference on Learning Representations (ICLR 2019). DOI: 10.48550/arXiv.1809.04281

[5] Nagathil Anil, Martin Rainer. (2013). Evaluation of spectral transforms for music signal analysis. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

[6] Bello, J. P., Chew, E., & Turnbull, D. (2008). ISMIR 2008: Proceedings of the 9th International Conference of Music Information Retrieval.

[7] Li, F., Guo, S., Han, F., Zhao, J., & Shen, F. (2024). Multi-scale dilated convolution network for long - term time series forecasting. arXiv preprint arXiv: 2405.05499.

[8] Kaiser, L., Gomez, A. N., & Chollet, F. (2017). Depthwise separable convolutions for neural machine translation. arXiv preprint arXiv: 1706.03059.

[9] Aguilera-Martos, I., Herrera - Poyatos, A., Luengo, J., & Herrera, F. (2024). Local attention mechanism: Boosting the Transformer architecture for long-sequence time series forecasting. arXiv preprint arXiv: 2410.03805.

[10] Ho, J., & Kalchbrenner, N. (2019). Axial attention in multidimensional transformers. arXiv preprint arXiv: 1912.12180.

[11] Zeng, M., Tan, X., Wang, R., Ju, Z. Q., Qin, T., & Liu, T. Y. (2021). MusicBERT: Symbolic music understanding with large - scale pre-training. Findings of the Association for Computational Linguistics: ACL - IJCNLP 2021, 791 - 800.

[12] Cai, Y., Liu, Y., Zhang, Z., & Shi, J. Q. (2023). CLAP: Isolating content from style through contrastive learning with augmented prompts. arXiv preprint arXiv: 2311.16445.

[13] Sharma, G., Srinivasamurthy, A., Gulati, S., & Serra, X. (2022). ACRNN: Attention-based Convolutional Recurrent Neural Network for Indian Classical Music Classification. Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR). DOI: 10.5281/zenodo.7316143

DOI: 10.54254/2755-2721/2025.BJ25786

[14] Saudi, A. S., Khalil, M. I., & Abbas, H. M. (2019). Improved features and dynamic stream weight adaption for robust audio-visual speech recognition framework. Digital Signal Processing, 89, 17-29.
[15] Yuan Gong, Yu-An Chung, James Glass. (2021). AST: Audio Spectrogram Transformer Interspeech. DOI: 10.21437/Interspeech.2021-698
[16] Jouppi, N., Kurian, G., Li, S., Ma, P., Nagarajan, R., Nai, L., ... & Patterson, D. A. (2023, June). Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In Proceedings of the 50th annual international symposium on computer architecture (pp. 1-14).
[17] Park, J. H., & cho, k. (2020). Large-scale self-supervised learning for music representation. In International Conference on Learning Representations (ICLR).
[18] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning (ICML) (pp. 1126 - 1135). PMLR.
[19] Tzanetakis, G., & Cook, P. R. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5), 293-302.
[20] Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X. Z., & Wei, F. (2023). BEATs: Audio pre-training with acoustic tokenizers. Proceedings of the 40th International Conference on Machine Learning, 202, 5178 - 5193.
[21] Elizalde, B., Deshmukh, S., Al Ismail, M., & Wang, H. (2022). CLAP: Learning audio concepts from natural language supervision. arXiv preprint arXiv: 2206.04769.