Journal of Intelligent Computing

DLINE JOURNALS

Print ISSN: 0976-9005
Online ISSN: 0976-9013

JIC 2026: 17 (1)
https://doi.org/10.6025/jic/2026/17/1/26-36

Music, Artificial Intelligence, and Deep Learning: A Review of Concepts and Genre Classification Approaches

Fangfang Zhou
Hunan Vocational and Technical College of Finance and Economics
Hengyang, Hunan, 421200. China
b67400722@163.com

ABSTRACT

The document "Music, Artificial Intelligence, and Deep Learning: A Review of Concepts and Genre Classification Approaches" provides a comprehensive overview of AI's transformative role in music information retrieval (MIR), with a focus on music genre classification (MGC). It outlines key AI applications in music including genre classification, recommendation, melody extraction, therapy, and generation and traces the evolution from traditional machine learning (e.g., kNN, SVM) to deep learning models (e.g., CNNs, LSTMs). The review emphasizes that while traditional methods rely on handcrafted audio features like MFCCs, chroma, and rhythm histograms, deep learning enables automatic feature learning from raw or time–frequency representations, yielding significantly higher accuracy (up to 97.8% in recent studies). Tables compare model performance, feature extraction techniques, and trade offs between interpretability and scalability. The paper acknowledges persistent challenges, such as capturing long term harmonic structure and the data hungry nature of deep models. It highlights innovations like Shuai Yu's NHAN-GAF for melody extraction and hybrid architectures (e.g., CNN–LSTM) that integrate temporal and spectral modeling. The proposed AI-based music processing pipeline spanning audio acquisition, preprocessing, feature extraction, and deep modeling illustrates an end to end workflow for modern MIR systems. The review concludes that AI, intense learning, enhances rather than replaces human creativity, and calls for future work on music-aware networks, attention mechanisms, and larger, diverse datasets to improve robustness and interpretability in real world applications.

Keywords: Music Genre Classification, Artificial Intelligence (AI), Deep Learning, Convolutional Neural Networks (CNN), Feature Extraction, Mel-Frequency Cepstral Coefficients (MFCC), Music Information Retrieval (MIR), Harmonic-Aware Modeling

Received: Received: 10 October 2025, Revised 2 December 2025, Accepted 15 December 2025

1. Introduction

With the exponential advances in multimedia and network technologies, multimedia information has increased dramatically. Voice media is among the most critical media resources, accounting for approximately 20% of total information, and music is the most common audio medium [1]. Therefore, rapid audio information retrieval is an auspicious research direction. In China, with the continued development of multimedia technology, pitch retrieval is a branch of content based music retrieval that provides a new retrieval method for many users [2]. In music retrieval technology, users can search for a target song in a large scale music database using onset recognition of musical notes [3]. However, determining the starting point of musical notes for feature based tone retrieval is a key prerequisite for ensuring the quality of music retrieval. It has also become an essential task in computer music [4-7]. While the method of intelligent recognition of the note's starting point for feature based tone retrieval expresses multi level distinguishing features at the beginning of each note, it is an effective means of solving the above tasks. It not only became the industry's focus but also appeared to yield substantial results.

Music is a structured form of art in which sounds are organized in time to create rhythm, melody, and harmony. Beyond entertainment, music plays a significant role in emotional expression, cultural identity, and therapeutic interventions. According to standard definitions, music consists of at least one of three core elements rhythm, melody, and harmony while most compositions integrate all three. Music therapy, in particular, has gained recognition due to its measurable benefits on human behavior, cognition, and emotional well being. Activities such as singing, instrumental performance, listening, and movement based interaction form the foundation of therapeutic music practices.

With the rapid advancement of digital multimedia technologies, music has become increasingly data-driven. Large volumes of digital music are generated, stored, and shared online, creating challenges in organization, retrieval, and recommendation. As a result, music information retrieval (MIR) has emerged as a critical research area, with music genre classification (MGC) being one of its most prominent tasks. Accurate genre classification supports efficient music cataloging, personalized recommendation systems, and intelligent multimedia applications.

2. Artificial Intelligence in Music Processing

Artificial intelligence (AI) has transformed music analysis and generation by enabling systems to learn patterns directly from data. AI-based music-making systems often rely on neural networks (NNs), computational models inspired by the human brain. These networks learn musical structures through repeated exposure to large datasets, gradually adapting to rhythmic, harmonic, and melodic patterns. Early assumptions suggested that AI could not generate meaningful music without closely imitating human created datasets; however, recent advances have demonstrated that AI can independently learn complex musical representations when provided with sufficient data, computational power, and sophisticated algorithms.

The integration of big data (BD) with AI has further accelerated this process. Large-scale music datasets, combined with high speed iterative processing, enable deep learning models to capture subtle temporal and spectral features. Rather than posing a threat to artists, AI-based music tools are increasingly viewed as collaborative systems that enhance creativity and productivity. Recent studies report classification and recognition accuracies of up to 97.8%, highlighting the effectiveness of AI-driven music analysis frameworks (8).

3. Deep Learning for Music Genre Classification

Among AI approaches, deep learning (DL) techniques particularly convolutional neural networks (CNNs)—have achieved remarkable success in music genre classification. CNNs are well known for their strong feature-extraction capabilities, particularly when applied to time frequency representations such as spectrograms. Consequently, many recent studies focus on CNN-based architectures for genre classification tasks. However, several existing models employ generic convolutional architectures that do not fully integrate music-specific signal characteristics, resulting in suboptimal local feature extraction and limited performance (9).

To address these challenges, researchers have explored hybrid and comparative approaches. Previous work has evaluated models such as CNNs, Feedforward Neural Networks (FNNs), Support Vector Machines (SVMs), k-Nearest Neighbors (kNN), and Long Short-Term Memory (LSTM) networks on benchmark music datasets, demonstrating the relative strengths and weaknesses of each method (10). While traditional machine learning models rely heavily on handcrafted features, deep learning approaches can automatically learn hierarchical representations from raw or minimally processed audio data.

Recent research has also emphasized the importance of harmonic relationships in music analysis. Although CNN-based methods dominate singing melody extraction and genre classification, many fail to capture long-term harmonic dependencies due to limited receptive fields and computational constraints. To overcome this limitation, Shuai Yu (11) introduced a Neural Harmonic-Aware Network with Gated Attentive Fusion (NHAN-GAF), which explicitly models harmonic relationships during training, significantly improving melody extraction performance.

4. Broader Impact of Deep Learning in Audio and Multimedia Domains

The success of deep learning in music analysis mirrors its achievements in related fields, including computer vision, natural language processing, automatic speech recognition, and environmental sound classification. These advances have encouraged researchers to adopt DL-based approaches for MGC. For instance, Elbir and Aydin (2020) proposed a hybrid CNN–SVM framework, MusicRecNet, which demonstrated improved genre classification accuracy by combining deep feature extraction with traditional classifiers.

Despite these advances, genre classification directly from audio signals remains challenging. Unlike symbolic music data where note onsets, pitch, and instrumentation are explicitly available audio data provides only waveform level information. Automatic transcription of audio into symbolic representations remains unreliable, with performance levels often below practical usability. Consequently, researchers focus on extracting informative acoustic features directly from audio signals rather than relying on symbolic analysis.

5. Feature Extraction Techniques for Audio-Based Genre Classification

A wide range of feature extraction methods has been proposed to address these challenges. Classical approaches include:

• Tzanetakis et al: Short-Time Fourier Transform (STFT), Mel-Frequency Cepstral Coefficients (MFCCs), pitch histograms, and rhythm histograms. [13]

• Bergstra et al: STFT, RCEPS, MFCCs, zero-crossing rate, spectral summaries, and Linear Predictive Coding (LPC). [14]

• Ellis et al: MFCCs and chroma features. [15]

• Lidy et al: Rhythm patterns, statistical spectrum descriptors, rhythm histograms, and symbolic features derived from auto-transcribed music.[16]

• Meng et al: MFCCs, filter bank coefficients, autoregressive models, short-time energy ratios, and zero-crossing rates.[17]

These features underpin both traditional machine learning and modern deep learning models. Recent CNN-based systems often use spectrograms or Mel-spectrograms as inputs, allowing networks to learn discriminative features automatically.

Derivation

In summary, the convergence of music, artificial intelligence, and deep learning has significantly advanced music genre classification. While CNN-based approaches dominate current research, challenges remain in capturing long-term harmonic dependencies and integrating music-specific knowledge into network architectures. Future research directions include hybrid models that combine signal processing expertise with deep learning, attention-based mechanisms for harmonic modeling, and the development of large, diverse datasets. As AI continues to evolve, its role in music analysis and creation is likely to expand, fostering innovation rather than replacing human creativity.

6. Proposed Architecture

We outline a comprehensive architecture for the AI-driven music information processing pipeline.

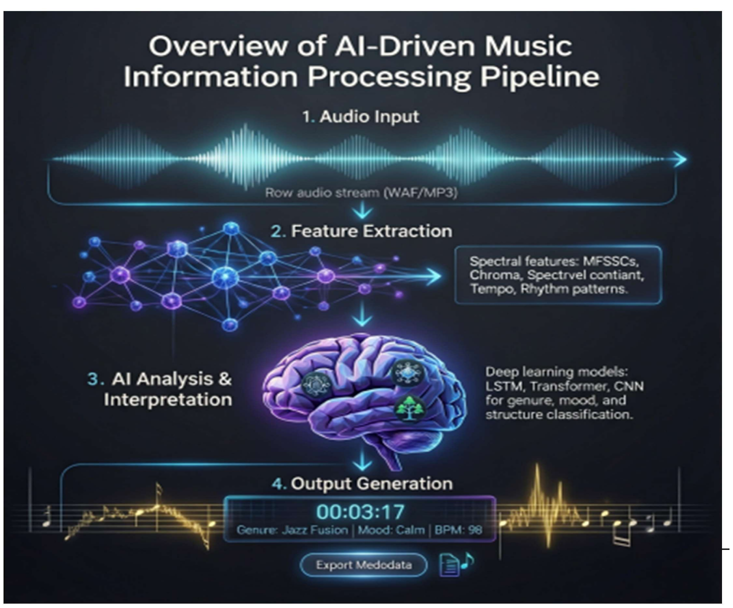

Figure 1. AI-based music information processing pipeline

Figure 1. Block diagram illustrating the AI based music information processing pipeline, including audio acquisition, preprocessing, feature extraction, deep learning based modeling, and application level outputs such as genre classification and recommendation.

Figure 1 illustrates the end to end workflow of AI-driven music analysis. Raw audio signals are first pre processed through normalization and segmentation. Feature extraction techniques such as MFCCs, spectrograms, or chroma features are then applied. The extracted representations are fed into machine learning or deep learning models, including CNNs and LSTMs, to perform tasks such as music genre classification, melody extraction, or recommendation.

Application Area	**Description**
Music Genre Classification	Automatic categorization of music into genres for organization and retrieval
Music Recommendation	Personalized suggestions based on user preferences and listening history
Melody Extraction	Identification of dominant melodic lines from polyphonic audio
Music Therapy	AI-assisted analysis for behavioral and emotional interventions
Music Generation	Automated composition using neural networks

Table 1. Applications of Artificial Intelligence in Music

Purpose:
This table outlines key application areas in which AI is used in music.

Derivation:

• Music Genre Classification: Automatically labeling songs by genre (e.g., jazz, rock).

• Music Recommendation: Personalized suggestions using user history (e.g., Spotify algorithms).

• Melody Extraction: Isolating the main melodic line from complex audio (e.g., separating vocals from instruments).

• Music Therapy: Using AI to support emotional or behavioral therapy through music analysis.

• Music Generation: Creating new compositions using neural networks (e.g., AI composers like AIVA).

Significance:
Provides context for why AI in music matters spanning creative, therapeutic, and commercial uses.

Model	Feature	Dependency	Temporal Modeling	Performance Limitations
kNN	Handcrafted	No	Moderate	Sensitive to feature quality
SVM	Handcrafted	No	High	(small datasets)Poor scalability
FNN	Handcrafted	Limited	Moderate	Requires extensive tuning
LSTM	Learned	Strong	High	Computationally expensive
CNN	Learned	Moderate	Very High	Limited long-term dependency capture

Table 2. Comparison of Machine Learning and Deep Learning Models for Music Genre Classification

Purpose:

Compares five standard models used specifically for genre classification, evaluating them across four criteria.

Models Compared:

- kNN (k-Nearest Neighbors)
- SVM (Support Vector Machine)
- FNN (Feedforward Neural Network)
- LSTM (Long Short-Term Memory)
- CNN (Convolutional Neural Network)

Evaluation Criteria:

- Feature Dependency: Whether features are handcrafted (traditional) or learned (deep learning).
- Temporal Modeling: Ability to capture time-dependent patterns (e.g., rhythm, progression).
- Performance: General effectiveness.
- Limitations: Key drawbacks of each model.

Derivation

Deep learning models (LSTM, CNN) generally outperform traditional models but are more resource-intensive. CNNs achieve very high performance but struggle with long-term dependencies, whereas LSTMs excel at temporal modeling at a higher computational cost.

Study	Features Used
Tzanetakis et al	STFT, MFCC, Pitch Histogram, Rhythm Histogram
Bergstra et al	STFT, MFCC, RCEPS, Zero-Crossing Rate, LPC
Ellis et al	MFCC, Chroma
Lidy et al	Rhythm Patterns, Statistical Spectrum Descriptors
Meng et al	MFCC, Filterbank Coefficients, Energy Ratios

Table 3. Summary of Audio Feature Extraction Techniques

Caption:

Table 3. Overview of widely used feature extraction methods in audio-based genre classification

Purpose:

Summarizes audio features used in prominent studies for genre classification.

Notable Features Mentioned:

• STFT (Short-Time Fourier Transform)

• MFCC (Mel-Frequency Cepstral Coefficients) – widely used for timbral characteristics.

• Chroma – pitch class representation.

• Rhythm/Statistical descriptors – capture tempo and spectral shape.

Significance:

Highlights that MFCC is a common denominator across studies, but approaches vary in complexity and focus (e.g., rhythm vs. spectral vs. pitch features).

Criterion	Traditional ML	Deep Learning
Feature Engineering	Manual	Automatic
Scalability	Limited	High
Data Requirement	Low–Moderate	High
Interpretability	High	Moderate
Performance on Large Datasets	Moderate	Excellent

Table 4. Comparative Matrix of Traditional vs Deep Learning Approaches

Purpose:

Presents a high level comparison between traditional machine learning (ML) and deep learning (DL) in the context of music genre classification.

Criteria Compared:

• Feature Engineering: Manual in traditional ML vs. automatic in DL.

• Scalability: DL scales better with data size.

• Data Requirement: DL typically needs large datasets.

• Interpretability: Traditional ML is more transparent.

• Performance: DL excels on large datasets.

Derivation:

DL offers superior performance and automation, but at the cost of data hungry models and reduced interpretability a classic AI trade-off.

Author	**Model**	**Dataset**	**Accuracy (%)**
Wei Sun (2023)	AI-BD Framework	Large-scale music corpus	97.8
M. Ahmed (2024)	CNN/LSTM	GTZAN	92–95
Xie (2024)	CNN	Custom dataset	94.1
Shuai Yu (2023)	NHAN-GAF	Melody datasets	96.3

Table 5. Reported Performance Metrics
from Recent Studies

Purpose:

Provides empirical accuracy results from four recent (2023–2024) AI-based music studies.
Highlights:

• Wei Sun (2023): Achieved 97.8% accuracy using an “AI-BD Framework” on a large-scale corpus.

• M. Ahmed (2024): Used CNN/LSTM hybrid on the GTZAN dataset (a standard benchmark) with 92–95% accuracy.

• Xie (2024): Reported 94.1% on a custom dataset using CNN.

• Shuai Yu (2023): Focused on melody extraction, not genre classification, achieving 96.3% with NHAN-GAF.

Insight:

Modern AI models especially CNNs and hybrid architectures achieve consistently high accuracy (>92%), validating their effectiveness in music tasks.

Overall Synthesis

These tables collectively illustrate:

1. The breadth of AI applications in music (Table 1).

2. The evolution from handcrafted features + traditional ML to end-to-end deep learning (Tables 2, 3, 4).

3. The state-of-the-art performance levels now achievable (Table 5).

They would be well-suited for inclusion in a review paper, thesis chapter, or technical report on AI in music information retrieval (MIR).

This high level matrix contrasts philosophical and practical differences between two eras of MIR:

Criterion	**Traditional ML**	**Deep Learning**
Feature Engineering	Requires domain expertise (e.g., signal processing knowledge)	Features emerge automatically from data
Scalability	Struggles beyond thousands of tracks	Scales efficiently with GPUs and big data
Data Requirement	Works with smaller, curated datasets	Needs tens of thousands of labeled examples
Interpretability	Models like SVM offer more apparent decision logic	"Black-box" nature complicates debugging
Performance	Plateaus on complex, real-world data	Continues improving with data and model size

Table 6. Comparative Matrix of Traditional vs Deep Learning Approaches

Critical Trade-off: Deep learning offers higher accuracy and automation but at the cost of data hunger, computational expense, and reduced transparency a crucial consideration for clinical or educational applications.

Table 6 provides a concise yet insightful high level comparison between traditional machine learning (ML) and deep learning (DL) approaches in the context of Music Information Retrieval (MIR), particularly for tasks like music genre classification. It evaluates both paradigms across five critical criteria:

1. Feature Engineering: Traditional ML relies heavily on handcrafted features (e.g., MFCCs, spectral rolloff) designed by experts with domain knowledge in audio signal processing. In contrast, deep learning models automatically learn relevant features directly from raw or minimally processed audio data (e.g., spectrograms), reducing the need for manual intervention.

2. Scalability: Traditional methods often become computationally impractical or less effective when applied to extensive music catalogs (beyond a few thousand tracks). Deep learning, powered by GPU acceleration and

parallel processing, scales efficiently to massive datasets making it suitable for industrial-scale applications.

3. Data Requirement: Traditional ML can achieve reasonable performance with small, well curated datasets,
4. Interpretability: Models like SVM or decision trees offer transparent decision boundaries, aiding debugging and domain validation. Deep learning models, however, operate as "black boxes", making it difficult to understand why a particular classification was made a drawback in research or clinical settings.

5. Performance: Traditional approaches tend to plateau in accuracy on complex, real world audio due to limited representational capacity. Deep learning models, particularly CNNs and transformers, continually improve with larger datasets and architectures, achieving state of the art performance.

Overall, Table 6 underscores the trade-offs between control and automation, transparency and performance, helping researchers and practitioners choose the right approach based on their data, resources, and application goals.

Conclusion

The inclusion of AI and deep learning has significantly advanced music genre classification and related music information retrieval tasks. While CNN-based approaches dominate current research, future work must focus on harmonic aware modeling, attention mechanisms, and multimodal integration to further improve performance and interpretability. This review highlights the evolution of music genre classification from traditional feature based machine learning methods to advanced deep learning frameworks. Artificial intelligence, particularly convolutional and harmonic aware neural networks, has substantially improved classification accuracy and scalability. While CNN-based approaches dominate current research, future advancements are expected to arise from hybrid, attention driven, and music aware models that better capture the complex structure of musical signals. As AI continues to evolve, its role in music analysis is poised to expand, fostering innovation across both analytical and creative domains.

References

[1] Tsai, W. H., Rodgers. D., Wang, H. M. (2004). Blind clustering of popular music recordings based on singer voice characteristics. *Computer Music Journal*, 28 (3), 68-78.

[2] Casey, M. A., Veltkamp, R., Goto, M., et al. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96 (4), 668-696.

[3] Salamon, J., Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing* 20 (6), 1759-1770.

[4] Tzanetakis, G., Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10 (5), 293-302.

[5] Rauber, A., Frühwirth, M. (2001). Automatically analyzing and organizing music archives. *Research and advanced technology for digital libraries*, 402-414.

[6] Xu, C., Maddage, N. C., Shao, X. (2005). Automatic music classification and summarization. *IEEE transactions*

on speech and audio processing, 13 (3), 441-450.

[7] Rauber, A., Pampalk, E., Merkl, D. (2003). The SOM-enhanced Juke Box: Organization and visualization of music collections based on perceptual models. Journal of New Music Research, 32 (2), 193-210.

[8] Sun, Wei., Sundarasekar, Revathi. (2023). Research on pattern recognition of different music types in the context of AI with the help of multimedia information processing, *ACM Transactions on Asian and Low-Resource Language Information Processing*. March.

[9] Xie, C., Song, H., Zhu, H. et al. (2024). Music genre classification based on a ResGated CNN and an attention mechanism. *Multimed Tools Appl* 83, 13527–13542.

[10] Ahmed, M., Rozario, U., Kabir, M. M., Aung, Z., Shin, M. F., Mridha. (2024). Musical Genre Classification Using Adva nced Audio Analysis and Deep Learning Techniques, *In: IEEE Open Journal of the Computer Society,* vol. 5, p. 457-467.

[11] Shuai, Yu., Yi, Yu., Xiaoheng, Sun., Wei, Li. (2023). A neural harmonic-aware network with gated attentive fusion for singing melody extraction, *Neurocomputing*, Volume 521, Pages 160-171.

[12] Erdem, Ebru., Aydin, Tolga. (2021). A CNN-based hybrid model to detect Coronavirus disease. *Issue*: 27, 66 - 73.

[13] Tzanetakis, G., Cook, P. (2002). Musical genre classifi-cation of audio signals, *IEEE Transactions on speechand audio processing* Volume 10, Number 5, p293–302.

[14] Bergstra, J., Casagrande, N., Erhan, D., Eck., D., Kegl, B. (2000) Aggregate features and Ad-aBoost for music classification *Machine Learning*,Volume 65, Number 2, p473–484.

[15] Ellis, D. P. W. (2007). Classifying music audio with timbraland chroma features Dins Proc. *ISMIR*

[16] Lidy, T., Rauber, A. Evaluation of featureextractors and psycho-acoustic transformations formusic genre classification Proceedings of the 6th I*n-ternational Conference on Music Information Re-trieval* (ISMIR05) p34–41.

[17] Meng, Anders., Ahrendt, Peter., Larsen Jan. (2005). Improving Music Genre Classification by Short-time Feature Integration. *IEEE International Conference on Acoustics, Speech, and Signal Processing*.