SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(4), 1162.
https://doi.org/10.62617/mcb1162

Article

Research on multi-label classification model design in online teaching and learning for music scholars from a biomechanical perspective

Jianyu Shi[1,*], Zijun Shi[2]

[1] School of Normal, Xiangyang Polytechnic, Xiangyang 441050, China
[2] Faculty of Arts, Lingnan University, Hong Kong 999077, China
*** Corresponding author:** Jianyu Shi, 13995764766@163.com

CITATION

Shi J, Shi Z. Research on multi-label classification model design in online teaching and learning for music scholars from a biomechanical perspective. Molecular & Cellular Biomechanics. 2025; 22(4): 1162. https://doi.org/10.62617/mcb1162

ARTICLE INFO

Received: 10 December 2024
Accepted: 31 December 2024
Available online: 13 March 2025

Abstract: Music serves as a vital medium for emotional expression and cultural heritage, evolving significantly through advancements in digital education. This study introduces an integrated framework to enhance online music education via innovative music generation and genre classification techniques. Central to this research is the MGED (Music Generation and Education Development) framework, which utilizes Convolutional Neural Networks (CNNs) for feature extraction and Highway Networks for deep learning. By incorporating spatial attention mechanisms with Bidirectional Gated Recurrent Units (Bi-GRU), the framework generates high-fidelity spectrograms, while the Griffin-Lim algorithm ensures temporal coherence in the outputs. For genre classification, the study employs the CMBRU (Classification Model for Bi-GRU and Residual Units) framework, leveraging Mel-Frequency Cepstral Coefficients (MFCC) and multi-channel CNNs to achieve robust representation learning. This model effectively captures temporal dependencies, resulting in over 70% accuracy across five genres. Additionally, this research explores the design of a multi-label classification model tailored for online teaching and learning environments aimed at music scholars, viewed through a biomechanical lens. As online education becomes increasingly prevalent, the need for effective classification models that can handle multiple labels simultaneously is critical, particularly in music education, where diverse skills and knowledge areas intersect. The study employs biomechanical principles to analyze the physical aspects of music performance and learning, integrating these insights into the classification model. This approach not only enhances the educational experience for music scholars but also contributes to the broader field of online music education, paving the way for future research and practical applications.

Keywords: music genre; encoder and decoder; deep learning; online learning; biomechanical

1. Introduction

As a fundamental pillar of human culture, music has played an essential role in society since ancient times. Beyond merely being an art form, music acts as a powerful vehicle for emotional expression, creating connections between thoughts, feelings, and the human spirit through its elements of melody, rhythm, and timbre. Various musical styles and genres, each influenced by distinct cultural backgrounds and historical developments, reflect a wide range of aesthetic values. For example, classical music is esteemed for its intricate structure and deep emotional resonance; jazz embraces individuality and spontaneity through improvisation, while pop music attracts audiences with its catchy melodies and lively rhythms. This rich tapestry of musical diversity not only enriches the human experience but also plays a subtle role in shaping emotional regulation, cognitive growth, and social interactions [1].

From a biomechanical perspective, music education plays a vital role in cultivating the physical skills essential for effective musical performance. By understanding the mechanics of movement, musicians can significantly enhance their technical abilities, optimize their posture, and reduce the risk of injuries that often arise from repetitive motions or improper techniques.

This perspective underscores the critical importance of body awareness and coordination in executing musical pieces with precision and expressiveness. Musicians who develop a keen sense of their body's movements are better equipped to interpret music nuances, maintain consistent tone quality, and execute complex passages with ease.

Furthermore, incorporating biomechanics into music education can lead to tailored training programs that address individual needs, ultimately improving performance quality. Techniques such as breath control, finger dexterity, and muscle relaxation can be systematically taught, enabling musicians to achieve a higher level of artistry and physical comfort during performances. This holistic approach not only enhances musicality but also fosters a sustainable practice environment, allowing musicians to thrive in their craft over the long term.

As a pivotal avenue for cultivating musical literacy and artistic sensibility, music education exerts a profound influence on both individual growth and societal progress. During early childhood, music education fosters language acquisition, enhances memory, stimulates creativity, and nurtures an appreciation for aesthetics and emotional expression. For adolescents and adults, it transcends skill development, offering intellectual enlightenment, fostering cultural understanding, strengthening logical reasoning, and cultivating teamwork.

In recent years, shifting educational paradigms have directed music education towards enhanced accessibility, diversity, and personalization. The emergence of online education has further democratized learning, providing platforms that reach a wider audience. However, a significant challenge remains within practical teaching environments: How to effectively utilize technological advancements to improve music instruction, deepen musical understanding, and boost performance skills. This issue has become a central focus for both educators and researchers [2].

With the rapid advancement of AI, automated research in music generation and classification has achieved remarkable progress. AI not only introduces transformative tools for music creation and education but also unlocks unprecedented opportunities to understand, produce, and disseminate music with greater efficiency. In the realm of automatic music generation, deep learning-based generative models—such as Recurrent Neural Networks (RNN), Transformers, and Variational Autoencoders (VAE)—have gained widespread adoption. By analyzing the structures and patterns within vast repositories of musical data, these models can autonomously compose music that reflects specific styles and characteristics [3]. For instance, the Transformer architecture, renowned for its robust contextual modeling capabilities, excels at capturing intricate relationships in musical sequences, enabling the generation of harmonically cohesive melodies with high precision. In the context of online music education, automated music generation technologies facilitate personalized learning by providing students with tailored materials for targeted training and creative exploration. Similarly, music genre classification has undergone a transformative shift,

as traditional manual methods are increasingly supplanted by machine learning and deep learning techniques. Leveraging advanced feature extraction methods, such as MFCC and Short-Time Fourier Transform (STFT), in combination with models like Convolutional Neural Networks (CNN), RNN, and Bi-Directional Gated Recurrent Units (Bi-GRUs), these approaches enable highly efficient and accurate genre recognition. Such advancements hold significant implications for online music instruction and large-scale music data management, empowering educators to perform comprehensive music analysis while assisting students in understanding the structural and stylistic nuances of various genres. Moreover, genre classification technologies facilitate the development of personalized recommendation systems that suggest relevant musical pieces and learning resources tailored to learners' preferences and needs. In conclusion, music generation and classification technologies offer robust support for the innovative evolution of modern music education, enhancing its intelligence, efficiency, and accessibility. These advancements not only enrich the learning experience but also pave the way for a more dynamic and personalized educational process [4].

With the increasing complexity of data, deep learning has demonstrated remarkable advantages in processing unstructured, multimodal, and time-series data. Music, as an intricate carrier of unstructured and multimodal information, encompasses diverse elements such as audio waveforms, spectral features, and temporal sequences. This inherent complexity renders traditional methods inadequate for comprehensively capturing its underlying structures and patterns [5]. Deep learning, through its ability to automatically extract latent features via multi-layer neural networks, effectively addresses non-linear relationships and complex patterns within music data.

This study focuses on advancing music generation and classification within the domain of online music teaching, with the following key contributions:

1) Automatic Music Generation (MGED Framework): To address the task of automated music generation, this study introduces the MGED framework. Feature extraction is accomplished using a CNN combined with max-pooling operations, while the Highway Network is employed to capture high-level features of music data. A spatial attention mechanism is integrated to emphasize critical spectral regions, enhancing feature precision. Coupled with the Bi-GRU network, the framework achieves high-fidelity spectrogram generation and duration label prediction, ensuring structural and temporal coherence between the generated music and its original counterpart.
2) Music Genre Classification (CMBRU Framework): For the challenge of genre classification, this paper proposes the CMBRU framework, which leverages MFCC features and CNNs to perform deep feature extraction across multi-channel audio data. The fusion of one-dimensional projections with original features enhances the model's representational capacity, while the Bi-GRU network effectively models the temporal dynamics of music signals, enabling precise classification of diverse music genres. Experimental results validate the framework's robustness and superior performance in multi-genre classification tasks.
3) Comprehensive Experimental Validation: Utilizing the GTZAN public dataset, a

benchmark for online music generation and classification, this study conducts extensive experiments to evaluate the MGED and CMBRU frameworks. The results demonstrate that both frameworks surpass traditional methods in generating high-quality music and achieving accurate genre classification, particularly in the modeling of complex audio data and time-series relationships.

The remainder of this paper is organized as follows: Section 2 reviews related works on music generation and genre classification. Section 3 details the proposed frameworks, MGED and CMBRU. Section 4 presents the experimental setup and results. Finally, the conclusion is drawn in Section 5.

2. Related works

2.1. Music generation studies

With the growing demand for music, intelligent music generation technology has emerged as a prominent research focus within the field of generative modeling. Computer-generated music integrates computational techniques with music theory expertise to produce complete musical compositions. In this paradigm, the computer serves as an essential tool, employing model-based algorithms to generate audio sequences with discernible features. These sequences can be described using various parameters and visualized or played back through applications and software platforms.

DeepMind's WaveNet network [6] was a pioneering contribution, generating audio waveform data directly. Initially designed to produce realistic human-like speech signals in the Text-to-Speech (TTS) domain, WaveNet models sample raw audio data to extract musical features. This process is implemented through multilayer convolutional or dilated CNN, resulting in the generation of entirely new audio files. Somush et al. [7] introduced the SampleRNN model, which efficiently generates audio files that closely resemble original music. Similarly, Donahue et al. [8] proposed the WaveGAN network, an adversarial neural network adapted for intelligent music generation. Operating in unsupervised data environments, WaveGAN generates diverse audio samples, including sounds of birds, human voices, and musical instruments, demonstrating its versatility across multiple audio modalities. OpenAI [9] presented the MuseNet model, inspired by the GPT-2 framework for natural language processing (NLP) text generation. MuseNet utilizes unsupervised techniques to predict subsequent signal distributions within musical sequences, enabling the generation of music in the styles of various composers [10]. The model incorporates melodic, chordal, and rhythmic sequences into a multilayer RNN to produce monophonic melodies. Through multitask learning, it orchestrates these melodies into cohesive compositions. By integrating prior musical knowledge into the network's architecture, MuseNet ensures that each layer focuses on specific structural features of music. Experimental results indicate that MuseNet outperforms baseline models in generating stylistically coherent and high-quality music. Tsushima et al. [11] proposed an LSTM-based model combined with a graphical user interface (GUI), allowing users to interactively arrange and create their own music. This approach further highlights the role of deep learning in enhancing creativity and enabling user-driven customization in music generation. In summary, advancements in intelligent music generation—ranging from WaveNet's foundational waveform synthesis to MuseNet's composer-

style orchestration—demonstrate the transformative potential of deep learning. These technologies offer increasingly efficient and flexible methods for creating complex, high-quality musical compositions.

2.2. Classification study of music

In traditional music research, scholars have relied on statistical features such as interval differences, rhythm, and melody extracted from WAV format audio, utilizing accumulated domain knowledge. Metrics like the mean or standard deviation of these statistical features were often employed as input variables for experiments [12]. The advantage of manually identifying music attributes and signal features lies in their high expressiveness and differentiation, which simplifies algorithmic model design and enhances classification accuracy. However, this approach suffers from significant limitations, including time-intensive processes, heavy reliance on manual effort, and restricted applicability, making it challenging to extract deep, representative features of music efficiently.

Machine learning methods have since become the focal point of music genre classification research, offering automated and scalable solutions [13]. Sugianto et al. [14] proposed a CNN approach incorporating a voting mechanism, where the Mel spectrogram serves as input. Compared to MFCC, the Mel spectrogram captures richer information, including frequency, time, and amplitude characteristics. Utilizing a VGG16 network combined with a global max-pooling layer, their method achieved a classification accuracy of 71.87% on the GTZAN dataset. Yunus et al. [15] tackled genre classification using music acoustic features and clustering methods, leveraging self-encoders within the clustering process for feature extraction and classification. Liu et al. [16] introduced a multi-instance attention mechanism (MATT) based on multi-instance learning (MIL). MATT extracts low-dimensional audio features, such as chroma features, Mel-inverted spectral coefficients, and spectral centroid (SC), and encodes these features using a CRNN to predict music genres effectively. Prabhakar et al. [17] applied a transfer learning approach to the music genre classification task, successfully predicting 11 music genres, including rock, pop, country, folk, and metal. Hasib et al. [18] further advanced this work by proposing an active transfer learning method for music genre classification. Their approach identifies informative labels for a small subset of samples in an unlabeled dataset, achieving superior accuracy on large-scale, noisy datasets compared to traditional methods like Support Vector Machines (SVM) and random forests. These studies demonstrate the growing efficacy of deep learning and transfer learning techniques in music genre classification, enabling the extraction of robust features and improving performance across diverse and complex datasets.

The aforementioned research highlights that, with the continuous advancements in computational power and deep learning methodologies, significant progress has been made in generative research based on music signals. By leveraging diverse features of original music in combination with corresponding signal characteristics, generative networks have demonstrated remarkable effectiveness in accomplishing music generation tasks. Furthermore, by integrating generative music with original music, deep learning algorithms can be applied to classify and analyze relevant music

data with high precision. This approach holds considerable practical and analytical value for enhancing music teaching and facilitating student learning, offering innovative pathways for personalized education and comprehensive musical understanding.

3. Methodology

3.1. The feature extraction module

For sound signals, in addition to the traditional features described in Section 2.2, it is essential to process higher-dimensional features to extract critical information more effectively. To this end, this paper employs 1D Convolutional Neural Networks (1D-CNNs) and Mel spectral features to enhance the representation of music data. 1D-CNN is a deep learning model primarily utilized for feature extraction from one-dimensional time-series signals [19]. In the context of music processing, 1D-CNN can directly analyze waveform data from audio signals (e.g., PCM or WAV formats), enabling the capture of temporally localized features such as pitch, tempo, and dynamic variations. The operation of a typical 1D convolution process is mathematically expressed as:

$$y[i] = \sum_{k=0}^{K-1} x[i+k] \times w[k] + b \tag{1}$$

where $y[i]$ is the value of the output feature map, *x* is the corresponding audio signal, w represents the size, which needs to be determined according to the corresponding data sampling frequency and other characteristics, and *b* is the corresponding bias. After the convolution operation, the activation function and pooling layers are applied to finalize the feature extraction, enabling the transformation of the original audio waveform into high-level, abstract features. This process effectively enhances the representation of temporally localized patterns in the audio data.

In addition to convolutional features, this study also incorporates the Mel Spectrum—a spectral representation obtained by mapping the audio signal onto the Mel scale following a Fourier Transform. The Mel spectrum is particularly advantageous in audio feature extraction as it closely approximates human auditory perception of pitch [20]. This perceptual alignment makes it an essential feature for modeling the auditory nuances of music signals. The Mel spectrum is computed by applying a series of overlapping windows to the input audio signal *x*, where each window is represented as:

$$X(t,f) = \sum_{n=0}^{N-1} x[n] \times w[n-t] \times e^{-j2\pi fn} \tag{2}$$

where *x* is the original signal, *w* is the window signal and *f* is the frequency component. Based on obtaining the frequency domain signal *X* we can calculate the corresponding frequency component magnitude squared i.e. *P*(*f*) as follows:

$$P(f) = |X(t,f)|^2 \tag{3}$$

Building upon this foundation, the frequency is projected onto the Mel scale through a Mel filter bank, transforming the linear frequency distribution into a logarithmic one. The corresponding features are then extracted by computing the logarithm of the spectral energy filtered through the Mel scale.

3.2. Encoder and decoder module for music generation (MGED)

Upon completing the feature extraction process, the construction of the music generative network is undertaken. Given the widespread application of GANs in traditional generative models, this study adopts an analogous approach, incorporating encoder-decoder frameworks to facilitate deeper data analysis. The temporal relationships between identical notes play a pivotal role in spectrogram generation, necessitating the encoder to capture these temporal dependencies [21]. Here, the convolutional network introduced in Section 3.1, along with the corresponding Meier spectral features, serves as input to distinct components of the encoder and decoder. Building on this foundation, an RNN is employed as the primary structure of the encoder, thereby enabling the construction of the complete network. The architecture of the proposed network is illustrated in **Figure 1**.

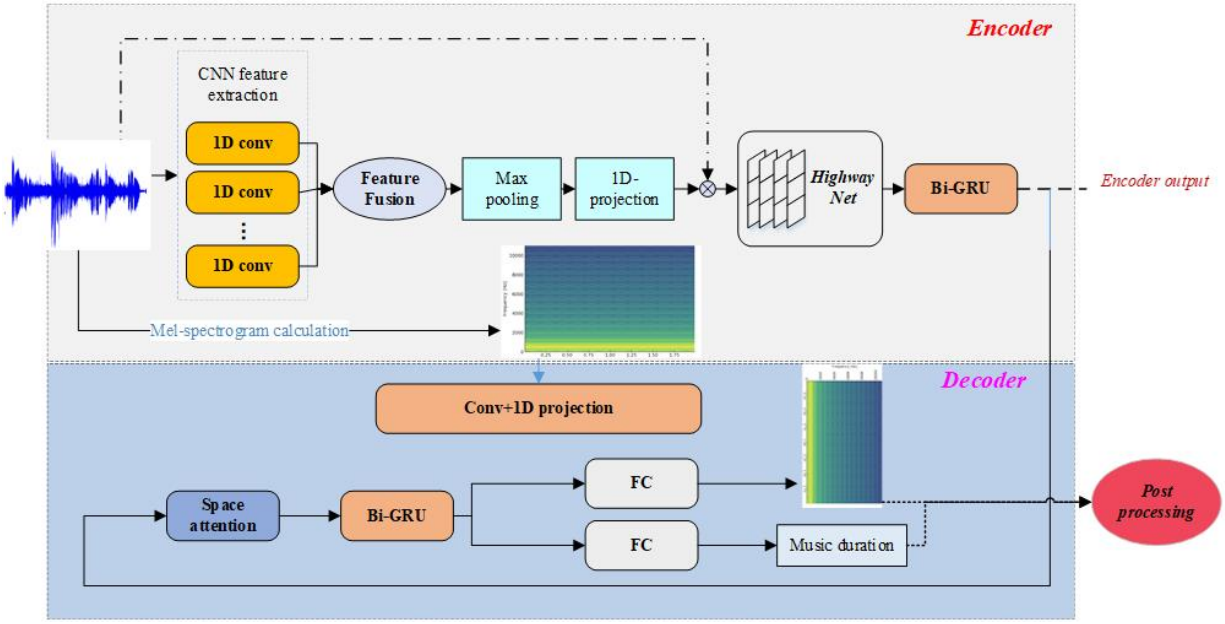

Figure 1. The framework for the MGED.

Within the MGED framework, which constitutes the music generation component of the music teaching process, the architecture is divided into two principal parts: the encoder and the decoder. In the encoder, the audio signal is initially processed through a convolutional network comprising n convolutional kernels, operating as defined in Equation (1). Following feature extraction, a max-pooling operation is applied to expand the model's receptive field while simultaneously reducing feature dimensionality. Subsequently, a 1D CNN is employed to project the multidimensional data onto a single dimension, ensuring consistency with the dimensionality of the input data.

Building upon this, the extracted feature data is fused with the original input data. Given that the features in musical data are predominantly low-level, Highway Networks are introduced to extract higher-level features by leveraging their capability to deepen network architectures without exacerbating training difficulty. Highway Networks, a deep learning architecture, are specifically designed to mitigate challenges such as gradient vanishing and gradient explosion that arise with increasing network depth. By incorporating a gating mechanism, these networks facilitate the smooth propagation of information across layers in a non-sequential, "shortcut" manner, enabling efficient training of deep architectures.

For the highway network, assuming that the input is x and the output is y, the formula of Highway network is as follows.

$$y = H(x, W_H) \times T(x, W_T) + x \times C(x, W_C) \tag{4}$$

H denotes the transform layer, which is used to change the input linearly; *T* denotes the transform gate, which is used to control the proportion of the output of the transform layer; *C* denotes the carry gate, which controls the proportion of the input passing through directly, usually complementary to *T*; and *W* is the weight parameter corresponding to each gate. Following the advanced feature extraction performed by this network, the Bi-GRU network is employed to further process the data, enabling an additional layer of feature extraction. Compared to traditional LSTM and RNN architectures, the Bi-GRU not only preserves hidden information within time series data but also facilitates the analysis of bidirectional dependencies.

The Bi-GRU primarily relies on two mechanisms—the reset gate and the update gate—to regulate the hidden states. These gates control the flow of information, allowing the network to efficiently manage temporal dependencies. The computational processes of the reset gate and update gate are defined in Equations (5) and (6).

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{5}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{6}$$

where the reset gate r_t controls how much information has been forgotten in the hidden state in the previous moment, while the update gate z_t can control how much the hidden state is updated in the current moment. After calculating the hidden state for forward and backward hidden states, respectively, we splice the two to form the current output h_t, and the whole process is shown in Equations (7)–(9):

$$\vec{h}_t = GRU(x_t, \vec{h}_{t-1}) \tag{7}$$

$$\overleftarrow{h}_t = GRU(x_t, \overleftarrow{h}_{t+1}) \tag{8}$$

$$h_t = [\vec{h}_t; \overleftarrow{h}_t] \tag{9}$$

Upon completing the design of the encoder, we constructed the corresponding decoder architecture, wherein the encoder's output is refined through a spatial attention mechanism. Subsequently, a Bi-GRU network layer is employed to facilitate the generation of new spectrograms and the labeling of music durations. The music duration outputs primarily serve to regulate the length of the generated spectrograms,

thereby ensuring temporal consistency between the generated data and the original input.

3.3. The establishment for the music generation and classification framework

After completing the encoding and decoding processes to generate the corresponding data, the spectrogram must be converted back into an audio waveform using a vocoder. In this study, the Griffin-Lim algorithm is employed as the vocoder [22]. The reconstruction of the audio waveform requires both the amplitude and phase spectra. While the linear spectrogram contains amplitude spectrum data, the Griffin-Lim algorithm estimates the phase information at each time step to reconstruct the waveform. The overall workflow of the algorithm is presented in Algorithm 1.

Algorithm 1 Overall Workflow

1: Input: Amplitude spectrum $|S(f,t)|$ (obtained from STFT)
2: Output: Reconstructed signal x
3: 1. Phase spectrum initiation $\phi_0(f,t)$
4: 2. for i=1:max iteration
5: 3. Complex spectrum calculation: $S_k(f,t) = |S(f,t)|e^{j\phi_k(f,t)}$
6: 4. ISTFT for Sk to obtain x. kx_k = ISTFT(S_k)
7: 5. Conducting STFT for xk to get $\tilde{S}_k(f,t) = \text{STFT}(x_k)$
8: 6. phase update $\phi_{k+1}(f,t) = \angle\tilde{S}_k(f,t)$
9: 7. compare -$\phi_{k+1}(f,t)\phi_k(f,t)$
10: 8. end

First, the target amplitude spectrum $|S|$ is input and the phase spectrum ϕ_0 is randomly initialised. Next, a complex spectrum is generated by combining the amplitude spectrum with the current phase spectrum $S_k = |S|e^{j\phi_k}$, and then an ISTFT is performed to restore the time-domain signal. The reconstructed time-domain signal is then subjected to another STFT to update the phase spectrum ϕ_{k+1}. This process is repeated until convergence or the maximum number of iterations is reached, and the reconstructed time-domain audio signal is finally output.

Upon completing the overall sound restoration, further decomposition of features and music style recognition becomes essential for applications in music education. To address this, we propose a music style label recognition algorithm based on multi-layer Bi-GRU and MFCC features. The overall structure of the proposed algorithm is illustrated in **Figure 2**.

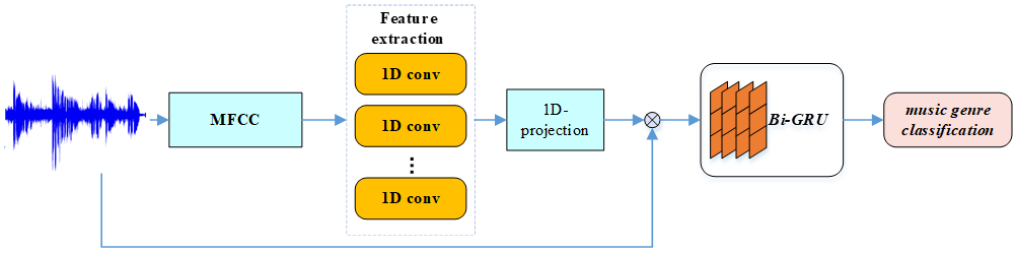

Figure 2. The framework for music genre classification.

The proposed framework, termed CMBRU, integrates CNN and MFCC features to achieve music genre classification through a Bi-GRU-based approach. Initially, MFCC feature extraction is performed on the generated music signal, which is then input into a 10-channel CNN network for feature analysis and extraction. The extracted features undergo a one-dimensional projection and are subsequently fused with the original features to establish a unified feature representation. Finally, the fused features are fed into the Bi-GRU module to accomplish music genre classification. Through this process, the study employs a coding-decoding structure to generate music while categorizing the generated outputs into specific genres. This approach facilitates further analysis, enabling more targeted music teaching and fostering aesthetic development.

4. Experiment result and analysis

4.1. Dataset and experiment setup

In this experiment, we address both music generation and genre classification tasks, utilizing two datasets: The GTZAN music genre library [23] and the repertoire performed during the Yamaha Piano Competition [24]. The GTZAN library comprises 10 distinct music genres, with 100 music clips per genre, totaling 1000 clips. Each clip varies in duration, typically ranging from 30 to 300 seconds. For this study, five genres were selected for analysis. Similarly, from the Yamaha competition database, the same five genres were extracted, and the dataset was augmented to form a unified basis for analysis.

To evaluate the performance of our proposed methods, we compared them with several state-of-the-art approaches from recent music classification studies, including the method proposed by Ghildiyal et al. [25], the framework by Sugianto et al. [14], the method of Prabhakar et al. [17], and the approach introduced by Yunus et al. [15].

Furthermore, to better examine the performance characteristics of individual modular components, we conducted an ablation study by evaluating the following variations: The CBRU model, excluding MFCC feature extraction; the MBRU model, omitting CNN features; and the BRU method, which employs GRU with raw data for classification alone. This comprehensive analysis allows for a deeper understanding of the contributions of each feature fusion component within the proposed model.

In the experimental analysis of the generated music, the models within the MGED framework were further divided into variants: GED, which employs Mel spectral features; GED-C, which excludes CNN features; and GED-A, which omits the Attention Module. These systematic comparisons and experiments allow for a thorough evaluation of each component's contribution to the overall generative performance. To address the validation challenges inherent in generative research, this study conducts a comparative analysis using the categorized music described in Section 3.3. Both objective evaluations and manual testing were employed, wherein objective metrics assess the label consistency between the generated music and the original music to validate the proposed models' effectiveness.

Upon confirming the experimental tasks, the next step involves establishing the experimental environment. Given the computational demands of the study, a high-

performance computing system was employed. The performance specifications of the selected system are presented in **Table 1**.

Table 1. The experiment environment information.

Item	Information
CPU	I5-14400F
GPUs	RTX 4080Ti
Language	Python 3.5.1
Framework	TensorFlow

4.2. Comparison of music tag classification effects

Following the introduction of the relevant datasets and methods, we proceeded to identify the various music genres within the adopted datasets and conducted a further analysis of their respective characteristics. The recognition results obtained using different methods are illustrated in **Figure 3**.

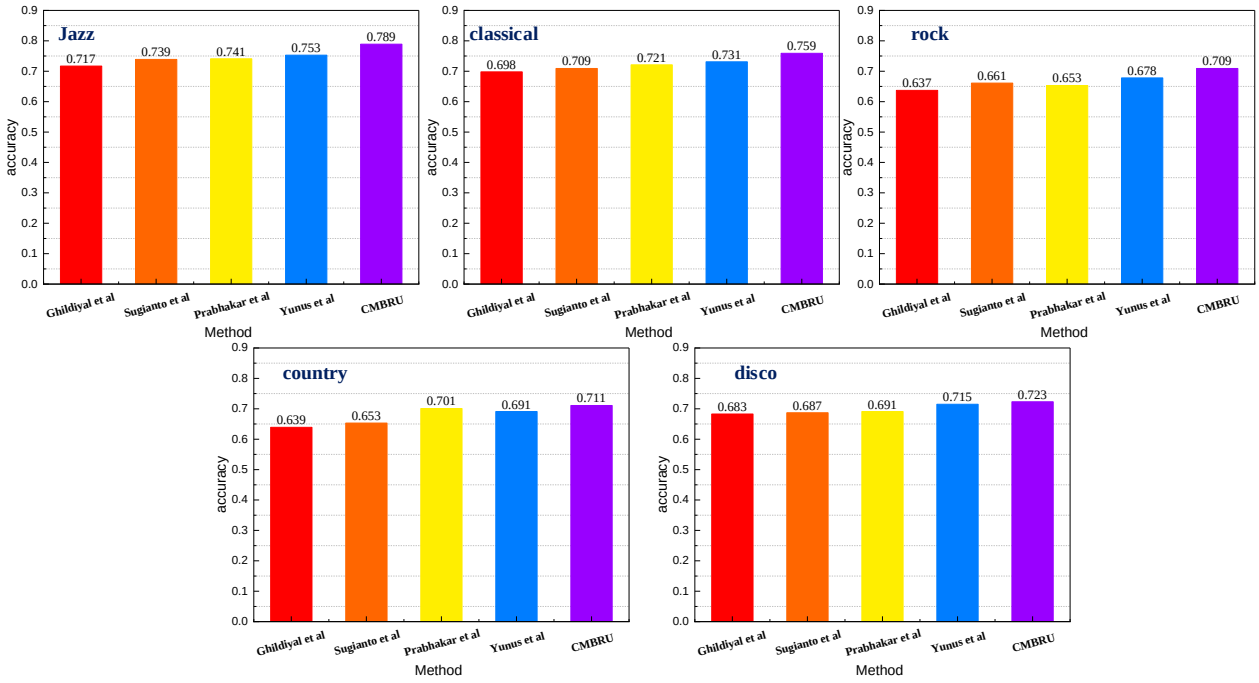

Figure 3. The method comparison results among different music genres.

As illustrated in **Figure 3**, the proposed CMBRU framework achieves genre identification performance exceeding 0.7 across different music genres, demonstrating a notable improvement over both existing studies and traditional methods. To further evaluate its robustness, we conducted a detailed analysis of the means and corresponding standard deviations for the five selected genres. The results of this analysis are presented in **Figure 4**.

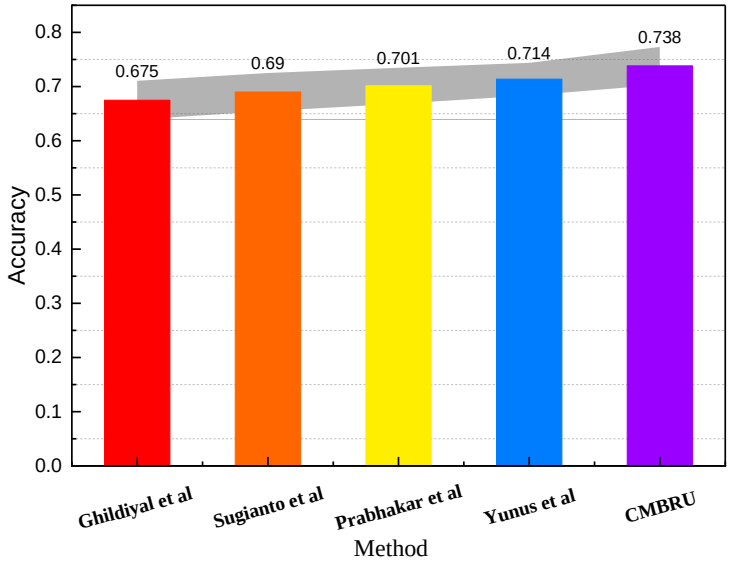

Figure 4. The statistical result for different methods.

As shown in **Figure 4**, the proposed CMBRU framework demonstrates superior stability, achieving an average recognition accuracy of 0.738 across the five music genres, with minimal variance fluctuations. This indicates that the framework maintains robust and consistent performance across different types of music. Following this comparison, model ablation experiments were conducted to further analyze the contributions of individual modules within the CMBRU architecture. The results of these ablation studies are presented in **Figure 5**.

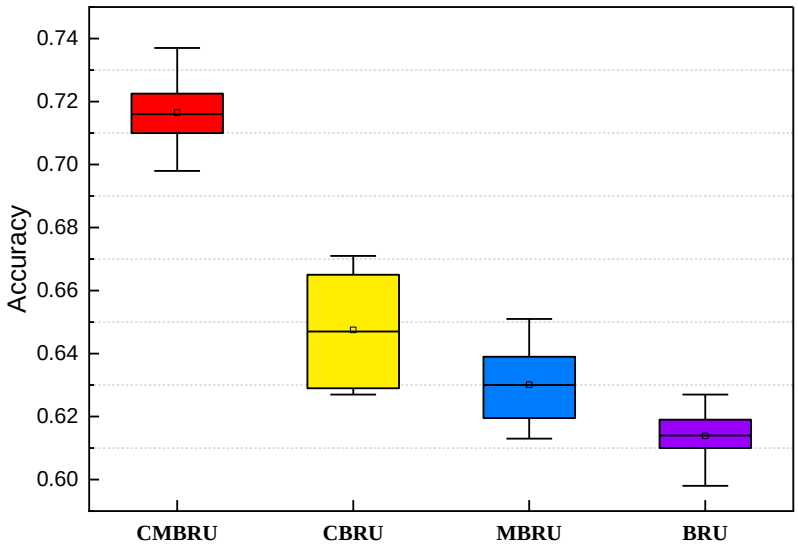

Figure 5. The Ablation experiment for different modules.

As observed in **Figure 5**, the CNN-extracted features play a pivotal role in music genre recognition. The MBRU model, which excludes these features, achieves an overall recognition accuracy of only about 0.65. In contrast, when CNN features are used without MFCC, the performance remains only slightly lower than that of the

complete CMBRU framework. This highlights that the CNN significantly enhances feature representation, making it the most crucial process in generative music analysis. In the subsequent section, we will further discuss the performance of the generative music module.

4.3. Generative music evaluation

Inspired by the ablation experiments conducted in Section 4.2 for music genre analysis, we extended our investigation to the task of music generation. Here, the performance of the generated data was evaluated with a focus on the contributions of individual modules. The reasonableness of the generative outputs was primarily assessed through manual annotation.

To this end, we designed two comparative experiments: one involving manual annotation and judgment of the generated music and the other using the CMBRU framework to classify the genres of the generated music. These results were compared with traditional approaches to enable genre classification and to observe the characteristics of the generated music. The consistency between the original data and the generated data, evaluated on a one-to-one basis, is presented in **Figure 6**.

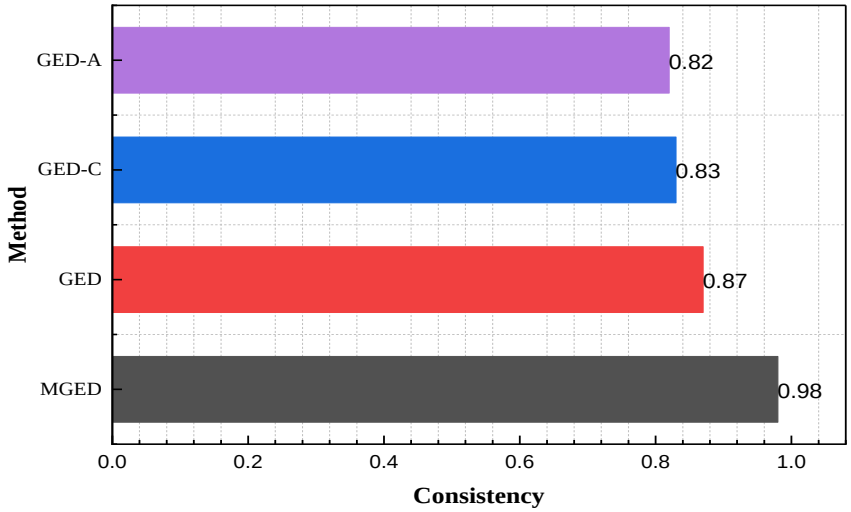

Figure 6. The human comparison result.

As illustrated in **Figure 6**, the MGED music generation framework achieves a satisfaction level of 0.98 under manual evaluation criteria, demonstrating excellent consistency and robustness of the generated music relative to the original music, even under varying degrees of human involvement.

In addition to the subjective classification, we further analyzed the genre correspondence between the original and generated data using the labeled CMBRU model. The results of this comparison are presented in **Figure 7**.

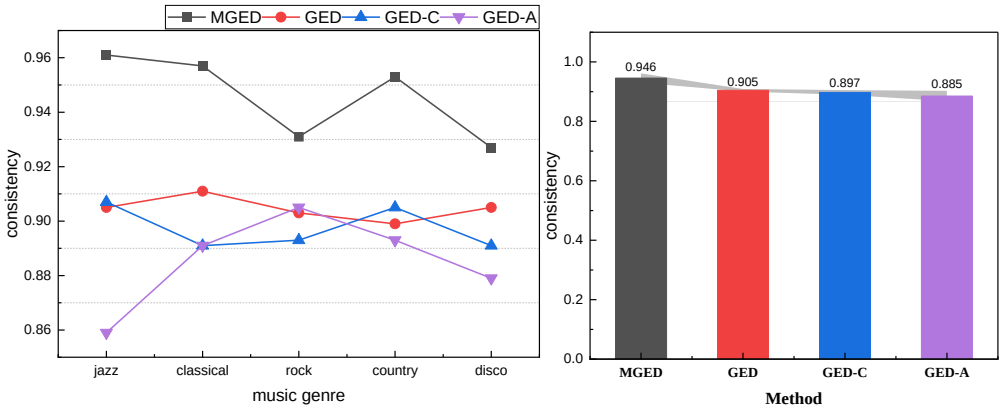

Figure 7. The generated music classification consistency.

As shown in **Figure 7**, the music generated by the MGED framework demonstrates high consistency in label classification, achieving exceptional performance, particularly in the jazz genre, where recognition accuracy exceeds 96%. Furthermore, in the overall statistical analysis, the MGED framework exhibits remarkable robustness on the generated data. Not only does the classification accuracy across all five genres approach 0.95, but the overall standard deviation remains minimal, highlighting the framework's reliability and effectiveness in successfully accomplishing the music generation task.

5. Discussion

In this paper, we investigate the automatic generation of audio data in music learning, specifically focusing on genre classification through labeled music. To address this, we propose the MGED framework in the music generation research section. In the encoder module, a convolutional network, combined with a max-pooling operation, effectively expands the model's receptive field while reducing feature dimensionality, ensuring efficient and accurate feature extraction. Additionally, a 1D-CNN is employed to project multidimensional data onto a one-dimensional space, maintaining consistency with the input data and enhancing its representational capacity. Feature fusion preserves critical information and enhances deep representation. To further optimize the depth of encoder-based feature extraction, the Highway network is introduced. Its gating mechanism effectively mitigates the issue of gradient vanishing during deep network training, enabling the extraction of more profound and advanced features without increasing training difficulty. This design surpasses traditional shallow neural networks or simple convolutional structures, capturing intricate and abstract music patterns with greater precision. The MGED framework employs a sophisticated decoder module where the integration of the spatial attention mechanism with a Bi-GRU network allows for the generation of new spectral data while maintaining temporal consistency with the original music. This is achieved by carefully regulating music duration labels, which serve as crucial temporal constraints. Such constraints substantially enhance the quality, coherence, and structural stability of the generated music, ensuring that it aligns closely with the intended artistic expression.

Unlike traditional music generation methods that often depend on rule-based approaches or single neural network structures—limiting their ability to preserve temporal continuity and multidimensional feature expressions—the MGED framework achieves significant improvements. By employing a dual-module encoding-decoding design and a highly efficient feature extraction mechanism, the framework markedly enhances both the accuracy and diversity of music generation.

This advancement is evidenced by superior results in both objective metric evaluations and manual comparison studies, demonstrating not only the robustness of the generated outputs but also the effectiveness of the proposed approach in various musical contexts. The ability to generate music that is not only coherent but also rich in texture and complexity positions the MGED framework as a leading solution in the field of automatic music generation, paving the way for innovative applications in online music teaching and beyond.

In the music genre classification task, the proposed CMBRU framework integrates CNN and MFCC features with a Bi-GRU network to deliver an efficient and accurate solution for music classification. MFCC features, as a classical method for audio feature extraction, effectively capture low-level and mid-level features of music signals from spectral information, making them well-suited for identifying timbre, pitch, and rhythmic characteristics that define music genres.

However, traditional approaches that rely solely on MFCC features and simple classifiers often fail to leverage deeper, more abstract feature representations. In the CMBRU framework, generated music signals are first processed through a multi-channel CNN network for deep feature extraction. This multi-channel design captures music features across various receptive fields, enhancing the model's capacity to perceive multi-scale information.

The integration of one-dimensional projection with the original features further fuses multi-level audio data representations, addressing the limitations of insufficient information expression found in traditional methods. Additionally, the Bi-GRU network serves as the classification module, leveraging its bidirectional recurrent structure to capture both forward and backward temporal dependencies within music signals. This enables a comprehensive understanding of the music's overall structure and genre characteristics. Compared to unidirectional RNNs or shallow classifiers, Bi-GRU offers superior contextual modeling capabilities, significantly enhancing the accuracy and robustness of the classification results.

6. Conclusion

In this paper, we conduct a comprehensive study on the automatic generation and genre classification of music within the context of online music teaching. We introduce an advanced framework designed to integrate Convolutional Neural Networks (CNNs), Highway Networks, Spatial Attention Mechanisms, and Bidirectional Gated Recurrent Units (Bi-GRUs). This innovative approach aims to address the shortcomings of traditional methods in the realms of complex audio feature extraction and time-series data modeling. The study initiates with a thorough pre-processing phase of audio data, which includes essential steps such as feature extraction and dimension mapping. This groundwork is crucial for establishing a solid

foundation for subsequent model training and music generation. The MGED framework leverages CNNs and Highway Networks to perform deep feature extraction, ultimately generating high-quality music spectra through the application of a spatial attention mechanism combined with Bi-GRU technology.

The generated spectra are subsequently converted into time-domain waveforms using the Griffin-Lim algorithm, ensuring the temporal and structural consistency of the resulting music. For music genre classification, the CMBRU framework is introduced, combining MFCC features and CNNs to extract multi-level audio features. The Bi-GRU further performs temporal modeling, enabling high-precision music genre classification. Experimental results on the classical GTZAN public dataset demonstrate the superior performance of the proposed frameworks. The CMBRU framework achieves a mean classification accuracy of 0.738 across five music genres, significantly outperforming traditional methods. In the generation task, the MGED framework achieves outstanding results, with generated music attaining over 0.9 agreement in both manual genre analysis and automatic genre classification under the CMBRU framework. These findings validate the effectiveness of the proposed frameworks in modeling complex audio data and improving feature representation, thereby advancing the techniques of automatic music generation and classification in online music teaching.

In future research, we aim to enhance the model's data processing capabilities by incorporating multimodal inputs, such as lyric text and sheet music information, to improve global feature modeling for both generation and classification tasks. Additionally, we will explore advanced optimization methods, including transformer architectures and adaptive attention mechanisms, to further enhance generation quality and classification accuracy. Moreover, we will investigate the adaptability of the proposed model across diverse music teaching scenarios, providing comprehensive technical support for the development of intelligent, dynamic music teaching systems. These efforts will establish a more robust theoretical and practical foundation for the continued evolution of digital music education.

Author contributions: Conceptualization, JS and ZS; methodology, JS and ZS; validation, JS and ZS; formal analysis, JS; data curation, ZS; writing—original draft preparation, ZS; writing—review and editing, JS and ZS; visualization, ZS; supervision, JS and ZS; project administration, JS. All authors have read and agreed to the published version of the manuscript.

Funding: This Project is supported by the 2023 Philosophy and Social Science Research Project of Hubei Provincial Department of Education: A Study on the Development of Music Appreciation Education in Chinese Ordinary Schools in the Past Century (1923-2022) Grant number 23G104.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments on this paper.

Availability of data and materials: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Kholjurayevich MN, Madaminovich FK. The role of music culture in inculcating the idea of national independence in the minds of the younger generation. ACADEMICIA: An International Multidisciplinary Research Journal. 2021; 11(5): 71–74. doi: 10.5958/2249-7137.2021.01351.3
2. Connolly R, D'Acierno P. Italian American Musical Culture and Its Contribution to American Music. The Italian American Heritage; 2021.
3. Ndou N, Ajoodha R, Jadhav A. Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches. In: Proceedings of the 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS); 2021.
4. Jamshidi F, Marghitu D, Chapman R. Developing an online music teaching and practicing platform via machine learning: a review paper. Cham: Springer International Publishing; 2021.
5. Xia Y, Xu F. Design and Application of Machine Learning-Based Evaluation for University Music Teaching. Mathematical Problems in Engineering. 2022; 2022: 1–10. doi: 10.1155/2022/4081478
6. Verma P, Chafe C. A Generative Model for Raw Audio Using Transformer Architectures. In: Proceedings of the 2021 24th International Conference on Digital Audio Effects (DAFx); 2021.
7. Mehri S, Kumar K, Gulrajani I, et al. SampleRNN: An unconditional end-to-end neural audio generation model. arXiv; 2019.
8. Donahue C, McAuley J, Puckette M. Synthesizing audio with generative adcersarial networks. In: Proceedings of the 10th International Conference on Learning Representations conference; 2019.
9. Subakan C, Ravanelli M, Cornell S, et al. Attention Is All You Need In Speech Separation. IEEE; 2021.
10. Zhu H, Liu Q, Yuan NJ, et al. XiaoIce Band: A melody and arrangement generation framework for pop music. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018.
11. Tsushima H, Nakamura E, Itoyama K, et al. Interactive arrangement of chords and melodies based on a tree-structured generative model. In: Proceedings of the International Society for Music Information Retrieval Conference; 2018.
12. Dineley J, Carr E, White LL, et al. Variability of speech timing features across repeated recordings: a comparison of open-source extraction techniques. In: Proceedings of the Interspeech 2024; 1–5 September 2024; Kos, Greece.
13. Cheng YH, Chang PC, Kuo CN. Convolutional Neural Networks Approach for Music Genre Classification. IEEE; 2020.
14. Sugianto S, Suyanto S. Voting-Based Music Genre Classification Using Melspectogram and Convolutional Neural Network. In: Proceedings of the 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI); 2019.
15. Atahan Y, Elbir A, Enes Keskin A, et al. Music Genre Classification Using Acoustic Features and Autoencoders. In: Proceedings of the 2021 Innovations in Intelligent Systems and Applications Conference (ASYU); 2021.
16. Liu X, Song S, Zhang M, et al. MATT. A Multiple-instance Attention Mechanism for Long-tail Music Genre Classification. In: Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2022.
17. Prabhakar SK, Lee SW. Holistic Approaches to Music Genre Classification using Efficient Transfer and Deep Learning Techniques. Expert Systems with Applications. 2023; 211: 118636. doi: 10.1016/j.eswa.2022.118636
18. Hasib KMd, Tanzim A, Shin J, et al. BMNet-5: A Novel Approach of Neural Network to Classify the Genre of Bengali Music Based on Audio Features. IEEE Access. 2022; 10: 108545–108563. doi: 10.1109/access.2022.3213818
19. Sharma G, Umapathy K, Krishnan S. Trends in audio signal feature extraction methods. Applied Acoustics. 2020; 158: 107020. doi: 10.1016/j.apacoust.2019.107020
20. Demir F, Turkoglu M, Aslan M, et al. A new pyramidal concatenated CNN approach for environmental sound classification. Applied Acoustics. 2020; 170: 107520. doi: 10.1016/j.apacoust.2020.107520
21. Shahriar S. GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. Displays. 2022; 73: 102237. doi: 10.1016/j.displa.2022.102237
22. Masuyama Y, Yatabe K, Koizumi Y, et al. Deep Griffin–Lim Iteration: Trainable Iterative Phase Reconstruction Using Neural Network. IEEE Journal of Selected Topics in Signal Processing. 2021; 15(1): 37–50. doi: 10.1109/jstsp.2020.3034486

23. SuriyaPrakash J, Kiran S. Obtain Better Accuracy Using Music Genre Classification Systemon GTZAN Dataset. In: Proceedings of the 2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon); 2022.
24. Edwards D, Dixon S, Benetos E, et al. A Data-Driven Analysis of Robust Automatic Piano Transcription. IEEE Signal Processing Letters. 2024; 31: 681–685. doi: 10.1109/lsp.2024.3363646
25. Ghildiyal A, Singh K, Sharma S. Music Genre Classification using Machine Learning. In: Proceedings of the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA); 2020.