A Comprehensive Dataset for Music Genre Classification Across 100 Genres: MuGen-100

Rihab Zouitni
School of Science and Engineering
Al Akhawayn University in Ifrane
Ifrane, Morocco
r.zouitni@aui.ma

Salma Samine
School of Science and Engineering
Al Akhawayn University in Ifrane
Ifrane, Morocco
s.samine@aui.ma

Yousra Chtouki
School of Science and Engineering
Al Akhawayn University in Ifrane
Ifrane, Morocco
y.chtouki@aui.ma

***Abstract*—Music genre classification is a fundamental task in Music Information Retrieval (MIR), which aids in organizing and retrieving large music collections. This paper introduces a comprehensive music genre dataset, the MuGen-100 Dataset, which consists of 100 music genres, with 100 song segments per genre, totaling 10,000 segments. The dataset was curated by collecting music from various online sources, predominantly YouTube playlists. Each song segment was extracted randomly, ensuring a diverse representation of each genre. The dataset is intended to serve as a benchmark for evaluating classification models in the context of genre recognition. Furthermore, it addresses the limitations of existing datasets by providing a larger, more diverse genre set, enhancing the robustness and generalization of machine learning and deep learning models. The dataset is available publicly for use in music genre classification research and applications.**

***Index Terms*—Music Genre Classification, Music Information Retrieval (MIR), Dataset, Deep Learning, Convolutional Neural Networks (CNNs), Transfer Learning, Data Augmentation, Audio Signal Processing.**

I. Introduction

Music genre classification plays a pivotal role in Music Information Retrieval (MIR), where the objective is to assign a label to a musical piece based on its genre. This task has gained significant attention due to its applications in music recommendation systems, content-based music retrieval, and automated playlist generation. The challenge in genre classification stems from the inherent complexity and diversity of music genres, which are often characterized by overlapping auditory features, such as rhythm, harmony, timbre, and instrumentation.

Traditional music genre classification methods have typically relied on handcrafted audio features like Mel-frequency cepstral coefficients (MFCCs), spectral contrast, and zero-crossing rate. However, these methods often struggle with complex musical structures and variations across different genres. Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have shown superior performance by learning hierarchical representations from raw audio data, such as spectrograms, without the need for explicit feature engineering. Nonetheless, the performance of these models heavily depends on the availability of high-quality, diverse datasets.

To address this challenge, we introduce the MuGen-100 Dataset, a comprehensive collection of 10,000 music segments covering 100 distinct genres. The dataset was curated from publicly available sources, ensuring a diverse representation of global musical traditions. It serves as a benchmark for evaluating genre classification models and enables researchers to explore new classification techniques and feature extraction methods.

The remainder of this paper is organized as follows: Section II reviews related work, including traditional and deep learning-based approaches to music genre classification. Section III provides an overview of existing music genre datasets and their limitations. Section IV presents the MuGen-100 Dataset, detailing its structure, data collection process, and preprocessing techniques. Section V discusses the challenges and considerations in dataset construction. Section VI outlines future research directions, and finally, Section VII concludes the paper with key takeaways.

II. Related Work

This section presents a review of relevant research on music genre classification. The discussion is structured into multiple sections, beginning with an overview of traditional machine learning approaches, followed by deep learning techniques, and concluding with a discussion of their limitations.

A. Traditional Machine Learning Approaches

Early methods for music genre classification relied on handcrafted audio features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral contrast, and rhythm patterns. Some studies focused on specific music traditions, utilizing MFCC-based feature extraction combined with classical machine learning models such as Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN) [1]. These approaches highlighted the challenges posed by genre complexity and overlapping musical characteristics.

Other research proposed classification frameworks incorporating rhythm-based and spectral features, demonstrating that feature engineering plays a crucial role in improving classification accuracy [2]. These studies emphasized the effectiveness of carefully selected audio features in enhancing model performance.

B. Deep Learning Approaches

With advancements in deep learning, convolutional neural networks (CNNs) have become the dominant approach for music genre classification. Some implementations have shown that CNN-based models, trained directly on spectrogram representations, significantly outperform traditional feature-based methods [3].

Recent studies have also explored transfer learning and data augmentation techniques to improve classification performance. Augmentation strategies involving reverberation and noise have been introduced to enhance model robustness in audio classification [4]. Additionally, research on **music source separation** has demonstrated that **transfer learning** can improve classification accuracy, especially for small datasets [5].

Comparative analyses between traditional machine learning and deep learning methods confirm that deep learning approaches, particularly CNNs, consistently outperform SVMs and decision trees [6]. Some studies have even reported that CNN-based models can achieve **human-level accuracy** in genre classification [7].

C. Limitations of Existing Approaches

Despite the success of deep learning models in music genre classification, several challenges remain. These include the computational expense of training deep neural networks, the need for large, labeled datasets, and the issue of genre overlap, which continues to impact classification accuracy. Additionally, many models struggle with handling diverse audio sources, such as noisy environments and varying recording qualities, making robust classification a persistent challenge.

III. Existing Music Genre Datasets

The availability of diverse and comprehensive datasets is crucial for the development of accurate and generalized music genre classification models. Numerous datasets have been created for the purpose of evaluating classification algorithms. The following provides a summary of several commonly used music genre datasets.

A. GTZAN Music Genre (2002)

The GTZAN dataset [8] is one of the most widely used benchmarks for music genre classification. It contains 1,000 tracks across 10 genres, with 100 tracks per genre. Each track is 30 seconds long and stored in mp3 format. The dataset is designed for single-label genre classification. Despite its popularity, there are several known issues, such as low-quality audio, artist repetition, and mislabeling, which can lead to overoptimistic classification results. Researchers often recommend using a cleaned and filtered version of the dataset to avoid these problems.

B. MagnaTagATune (2009)

MagnaTagATune [9] is a dataset designed for the music tagging problem, containing 5,405 tracks, each split into 29-second clips, with 230 artists and 188 tags. This dataset is useful for tasks that involve multi-label classification and tagging of music. However, it suffers from weak labeling, as the tags were generated through a gamified annotation process. Despite this, MagnaTagATune played a significant role in pioneering research in music tagging.

C. Million Song Dataset (2011)

The Million Song Dataset (MSD) [10] is the largest and most comprehensive music dataset available, containing approximately one million tracks. Although audio is not directly available, the dataset provides metadata and audio features, including last.fm tags, which are realistic and have been used in numerous deep learning-based music recommendation systems. MSD is primarily used for music tagging tasks and provides a large number of tags, but researchers need to be cautious of artist leakage and the availability of audio clips, as they are limited to 30-second previews.

D. FMA (2017)

The Free Music Archive (FMA) dataset [11] contains over 100,000 tracks of copyright-free indie music. It is a large-scale dataset ideal for genre classification and hierarchical classification tasks. The genre labels are chosen by the artists themselves from a pre-defined 163-genre hierarchy. While the dataset is large and diverse, the audio quality varies, and it is biased towards indie music, which may limit its applicability for other music classification tasks.

E. MTG-Jamendo (2019)

The MTG-Jamendo [12] dataset consists of 55,000 full tracks, each encoded at 320kbps mp3. It includes 195 tags spanning genres, instruments, and mood/themes. The dataset is particularly useful for tasks such as genre, instrument, and mood classification. The audio is high-quality and readily available, with predefined splits for training, validation, and testing. However, the genre distribution in MTG-Jamendo is skewed towards electronic music, which may affect the generalization of models trained on this dataset.

F. AudioSet (2017)

AudioSet [13] is a large-scale dataset designed for general audio understanding, with over 2.1 million audio clips, of which 1 million are related to music. The dataset provides strong temporal annotations for 10-second segments of YouTube videos. While the dataset includes a diverse range of music clips, it also contains low-quality audio, including live sessions and amateur recordings.

G. NSynth (2017)

NSynth [14] is a different dataset of 305,979 musical notes, each with a distinct pitch, timbre, and envelope. Although it is primarily designed for generative tasks rather than classification, it can still be used for proof-of-concept music classification experiments. The dataset is monophonic, sampled at 16 kHz with 4-second durations. However, it is not recommended for serious music genre classification tasks due to its focus on individual musical notes rather than complete songs.

H. Comparison of Music Genre Datasets

The following table summarizes the key characteristics of these datasets:

Dataset	Tracks	Genres	Audio Quality	Use Cases
GTZAN	1,000	10	Low (quality issues, artist repetition)	Single-label Genre Classification
Magna Tag ATune	5,405	188 Tags	Medium (weak labeling)	Music Tagging
Million Song Dataset	1M	522,366 Tags	High (metadata, no direct audio)	Tagging, Recommender Systems
FMA	100K+	163 Genres	Medium (indie music bias)	Genre, Hierarchical Classification
MTG-Jamendo	55K	195 Tags	High (full tracks)	Genre, Mood Classification
AudioSet	2.1M	Multiple Categories	Low (low-quality audio)	General Audio Understanding
NSynth	305,979 Notes	Single Notes	High (clean synthetic notes)	Music Generation, Proof-of-Concept

TABLE I
COMPARISON OF MUSIC GENRE DATASETS

IV. THE MUGEN-100 DATASET

This case study explores the dataset created for the task of music genre classification. The dataset consists of 100 music genres, each represented by 100 randomly segmented audio clips derived from songs obtained via YouTube. The process of dataset construction and the subsequent classification problem is elaborated below, detailing the genre categorization, data preparation, and challenges faced in creating and curating this dataset.

A. Dataset Overview

The dataset contains 100 distinct music genres, each with 100 audio samples. The genres have been carefully selected to cover a wide variety of global and regional styles, providing a rich and diverse corpus of musical data. Each genre is represented by three randomly chopped segments, each exactly 30 seconds in length, extracted from full-length songs sourced from YouTube. The genres span various cultural and musical traditions, allowing for a broad exploration of classification techniques. The 100 genres are as follows:

1) African and Middle Eastern Genres:

- Ahidous, Ahwach, Aita, Andalusi, Bikutsi, Kwaito, Gnawa, Mbalax, Wassoulou, Reggada, Soukous, Taarab, Baladi, Egyptian Shaabi, Hassani, Issawa, Moroccan Chaabi, Chgouri, Dakka Marrakchia, Makossa, Mbaqanga, Mbube, Taktouka Jabaliya, Griot, Highlife, Juju

This group includes a rich variety of traditional and contemporary music styles from Africa and the Middle East, often characterized by intricate rhythms, call-and-response singing, and deep cultural roots. Many of these genres are tied to rituals, storytelling, and dance. For example, **Gnawa** is a spiritual genre from Morocco with roots in African Sufi traditions, while **Mbalax** from Senegal blends traditional drumming with modern elements. **Kwaito** from South Africa incorporates house music with African sounds, and **Taarab** from East Africa fuses Arabic, Indian, and African influences.

2) European and Asian Genres:

- Classical Tarab, Balkan Brass, Alps, Gypsy, Nordic Folk, Slavic Folk, Chanson, Fado, Flamenco, Opera, Enka, Gamelan, Gugak, Bollypop, Carnatic, Ghazal, Semiclassical, Sufi, Iranian Classical, Turkish Classical, Turkish Folk, Didgeridoo, Haka, Pekinopera

This category features classical, folk, and popular music from Europe and Asia, reflecting a diversity of historical and cultural influences. **Fado** from Portugal is deeply emotional, expressing longing and nostalgia, while **Flamenco** from Spain is a passionate combination of song, dance, and guitar. In Asia, **Gamelan** from Indonesia showcases percussive, meditative music, and **Ghazal** from South Asia is a poetic form of classical singing. **Turkish Classical** and **Iranian Classical** music are rooted in ancient traditions, emphasizing microtonal melodies and expressive vocal techniques.

3) Western Popular Genres:

- Pop, City Pop, Jpop, Kpop, Old Cpop, Hip-Hop, R and B, Jazz, Blues, Funk, EDM, Metal, Rock, Rap, Reggae, Disco, Country, Salsa, Tango, Mariachi, Sertanejo, Cumbia, Rumba, Bachata, Son Cubano, Forro, Noteno, Bombay Plena, Samba

This section includes popular music styles that have gained international recognition, originating mainly in the Americas and Europe. **Pop, Hip-Hop, R&B, and Rock** dominate mainstream music, while **Jazz and Blues** have deeply influenced the development of modern music. Latin American genres like **Salsa, Tango, Mariachi, and Reggae** bring vibrant rhythms and danceable melodies. **EDM** represents the electronic music movement, while **Country and Funk** showcase storytelling and groove-based instrumentation.

B. Data Collection and Preprocessing

The process of curating this dataset involved multiple steps to ensure quality, accuracy, and consistency:

- **Song Selection**: A song was chosen for each genre from YouTube based on a detailed search of known tracks representative of the genre. The selection process involved not only identifying popular and culturally relevant tracks but also considering song structure, instrumentation, and genre characteristics. In addition to the YouTube-sourced music, we also used specific genres from a small MoroccanMusic dataset [15] and an Indian music dataset [16], which provided a rich representation of genres specific to those regions. Furthermore, we incorporated the genres from the well-known GTZAN dataset [8] to enhance the variety and robustness of our collection.
- **Audio Format Conversion**: The original audio format was converted to *.wav* files to standardize the input and ensure compatibility with machine learning models. The sampling rate was adjusted to ensure uniformity across the entire dataset.
- **Segmentation**: After downloading the song, the audio was chopped into three random segments, each exactly 30 seconds in length. This segmentation approach was designed to provide variety in terms of timing, allowing the model to learn features from different parts of the song.
- **Genre Labeling**: Each segment was placed into the corresponding genre folder, and filenames were annotated with the genre tag. This labeling ensured that each clip was correctly identified according to the genre it was extracted from.

C. Challenges and Considerations

The creation of this dataset was not without its challenges. Some of the key issues encountered during the process include:

- **Genre Ambiguity**: Certain genres, especially those with broad definitions such as *Pop* or *Classical*, have subgenres or overlap with other genres, which could result in ambiguity during classification. The genre definitions were kept as clear as possible, but ambiguities remain due to the evolving nature of music classification.
- **Audio Quality**: While YouTube provides a wealth of diverse content, the quality of the audio can vary significantly. Compression artifacts, background noise, and inconsistent volume levels can all impact the classification performance. These challenges were addressed during the preprocessing stage by applying noise filtering and normalization techniques.
- **Artist Repetition**: Given that songs from the same artists were often included in multiple genres, ensuring that the dataset did not inadvertently introduce biases from artist repetition was a significant concern. Steps were taken to minimize such repetition during the selection process.
- **Data Imbalance**: Some genres, such as *Pop* and *Rock*, are more popular than others, leading to potential class imbalances in the dataset. This imbalance was addressed by ensuring that all genres had an equal number of samples (100 segments per genre).

V. Discussion

The MuGen-100 Dataset introduced in this paper presents a significant advancement in the field of music genre classification. By encompassing 100 distinct music genres with 100 audio samples per genre, this dataset provides a diverse and comprehensive resource for researchers in the field of Music Information Retrieval (MIR). One of the primary challenges in genre classification is the overlapping characteristics of genres, particularly in cases where subgenres exhibit similar musical structures. Our dataset helps address this issue by including a wide spectrum of genres from various cultural and geographical backgrounds, ensuring a more balanced representation of global music styles.

Through our data collection process, we observed that some genres are underrepresented in existing datasets, making their classification more difficult. The inclusion of lesser-known genres, such as Ahidous, Wassoulou, and Gugak, presents an opportunity for researchers to explore and improve classification techniques for non-mainstream musical traditions. Additionally, the dataset provides high-quality audio segments extracted from various online sources, ensuring diversity in sound production and recording quality. This diversity introduces real-world challenges, such as handling variations in audio mixing, mastering, and recording environments, which must be considered when developing robust classification models.

VI. Future Work

As part of our ongoing research, we are currently evaluating the performance of Random Forest and CNN models on our dataset. Future work will involve an in-depth comparative analysis of these models, examining their classification accuracy, computational efficiency, and generalization across diverse musical genres. The results of this study will be presented as a continuation of this work, providing further insights into the efficacy of different classification techniques for music genre recognition.

Future research directions in music genre classification should emphasize enhancing model robustness, optimizing computational efficiency, and improving generalization to diverse datasets. Given the increasing complexity of deep learning architectures, a key focus should be on designing hybrid models that effectively integrate traditional machine learning techniques with advanced neural networks. Incorporating explainability methods, such as attention mechanisms, can facilitate a more interpretable classification process while improving model reliability.

Additionally, multi-modal learning approaches hold promise for refining genre classification. Future studies should explore the integration of audio features with metadata, lyrics, and contextual descriptors to enrich feature representations. Leveraging self-supervised and contrastive learning techniques can further enhance feature extraction, reducing reliance on large labeled datasets.

Expanding the dataset by incorporating additional musical attributes, such as tempo, instrumentation, and mode labels,

is crucial for enabling finer-grained classification. This would allow models to capture intricate genre-specific characteristics beyond standard categorical classification. Furthermore, developing benchmark tasks based on this dataset, including transfer learning applications and domain adaptation studies, would provide valuable insights into model performance under varying conditions.

Lastly, open-source contributions remain a priority. Future work should focus on providing reproducible frameworks, standardized evaluation metrics, and pre-trained models to facilitate further advancements in the Music Information Retrieval (MIR) community.

VII. Conclusion

This paper introduced the MuGen-100 Dataset, a comprehensive collection of 10,000 audio segments spanning 100 distinct music genres. By addressing the limitations of existing datasets, this resource enhances the diversity and robustness of music genre classification models. The dataset was carefully curated from publicly available sources, ensuring a wide representation of global musical traditions while maintaining a standardized format suitable for machine learning applications.

Through its extensive genre coverage and high-quality segmentation, the dataset provides a reliable benchmark for evaluating classification models. It also presents real-world challenges such as genre overlap, audio quality variations, and cultural diversity in music, making it a valuable tool for advancing research in Music Information Retrieval (MIR). By making this dataset publicly available, we aim to contribute to the development of more accurate, adaptable, and inclusive genre classification methodologies.

References

[1] B. Kumaraswamy, N. Kumari, T. Shukla, Swati, and K. Satyam, "Music genre classification for indian music genres," in *International Journal for Research in Applied Science and Engineering Technology*, vol. 9, 08 2021.

[2] J. Martins de Sousa, E. Torres Pereira, and L. Ribeiro Veloso, "A robust music genre classification approach for global and regional music datasets evaluation," in *2016 IEEE International Conference on Digital Signal Processing (DSP)*, 2016, pp. 109–113.

[3] S. Vishnupriya and K. Meenakshi, "Automatic music genre classification using convolution neural network," in *2018 International Conference on Computer Communication and Informatics (ICCCI)*, 2018, pp. 1–4.

[4] S. Huq, P. Xi, R. Goubran, J. J. Valdés, F. Knoefel, and J. R. Green, "Data augmentation using reverb and noise in deep learning implementation of cough classification," pp. 1–6, 2023.

[5] C. Garoufis, A. Zlatintsi, and P. Maragos, "Pre-training music classification models via music source separation," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, 2024, pp. 411–415.

[6] N. Ndou, R. Ajoodha, and A. Jadhav, "Music genre classification: A review of deep-learning and traditional machine-learning approaches," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1–6.

[7] M. Dong, "Convolutional neural network achieves human-level accuracy in music genre classification," *cornell university, arxiv)*, 2018.

[8] A. Olteanu, "Gtzan dataset: Music genre classification," https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification, 2020, [Accessed: Mar. 9, 2025].

[9] M. T. García, "Magnatagatune dataset," https://soundcloud.com/user-537934052/sets/magnatagatune, 2019, [Accessed: Mar. 9, 2025].

[10] ——, "Million song dataset," https://soundcloud.com/user-537934052/sets/million-song-dataset, 2019, [Accessed: Mar. 9, 2025].

[11] ——, "Free music archive dataset," https://soundcloud.com/user-537934052/sets/free-music-archive-dataset, 2019, [Accessed: Mar. 9, 2025].

[12] ——, "Mtg-jamendo dataset," https://soundcloud.com/user-537934052/sets/mtg-jamendo-dataset, 2019, [Accessed: Mar. 9, 2025].

[13] J. R. Fan, J. R. Green, and F. Knoefel, "Audioset: An ontology and collection of audio events," https://research.google.com/audioset/eval/music.html, 2017, [Accessed: Mar. 9, 2025].

[14] M. T. García, "Nsynth dataset," https://soundcloud.com/user-537934052/sets/nsynth-dataset, 2019, [Accessed: Mar. 9, 2025].

[15] R. Noureddine, "Moroccan music genre classification," https://www.kaggle.com/datasets/noureddineridanr96/moroccan-music-genre-classification, 2020, [Accessed: Mar. 9, 2025].

[16] W. Winchester, "Indian music genre dataset," https://www.kaggle.com/datasets/winchester19/indian-music-genre-dataset, 2020, [Accessed: Mar. 9, 2025].