Léa Runesha

Music Genre Classification Using Deep Learning: A Convolutional Neural Network Approach

Music genre classification plays a foundational and increasingly critical role in the domain of music information retrieval (MIR), particularly as the consumption of digital music continues to scale across global streaming platforms and multimedia applications. The exponential growth in digital music content has not only expanded the volume of audio data available to users, but also elevated the importance of intelligent indexing, retrieval, and recommendation systems that are capable of efficiently organizing and navigating large-scale music libraries. In such environments, accurate genre labeling serves as a cornerstone for content discovery, user personalization, and contextual metadata generation. Historically, genre classification has been performed through manual curation by musicologists, content editors, or platform-specific tagging experts. While effective in limited contexts, these traditional methods are resource-intensive, time-consuming, and inherently subjective, often leading to inconsistencies in labeling criteria and taxonomy interpretation. The lack of standardization and scalability in manual genre annotation has rendered it insufficient for meeting the demands of modern digital music ecosystems. In response to these limitations, the application of data-driven, machine learning–based solutions has gained considerable momentum. Among these, deep learning models—particularly convolutional neural networks (CNNs)—have emerged as a compelling approach for automating the genre classification task. CNNs excel at learning hierarchical feature representations from structured data, and when applied to audio inputs converted into spectrograms, they are capable of identifying salient frequency and temporal patterns indicative of specific musical genres. By reframing the problem of audio classification as a visual recognition task, CNNs capitalize on proven methodologies in computer vision and extend their utility into the auditory domain. This paradigm shift allows for scalable, objective, and reproducible genre classification pipelines that are not only efficient but also adaptable to a wide variety of musicological contexts and technological applications.

The goal of this project is to develop a robust deep learning-based framework capable of automatically classifying songs into distinct musical genres by analyzing visual features extracted from raw audio signals. Unlike traditional approaches that rely on manually defined heuristics or handcrafted features, this framework leverages data-driven representations to model the inherent structure of audio through deep convolutional architectures. At the core of this methodology is the transformation of waveform audio into Mel-spectrograms—a perceptually motivated representation of sound that captures frequency and temporal information in a two-dimensional format. This reframing of an audio task as a computer vision problem opens the door to applying state-of-the-art CNN techniques, which have revolutionized image recognition, to the domain of music classification.

This research not only advances the technical boundaries of music information retrieval (MIR) but also demonstrates the potential for wide-reaching practical applications. Commercial music streaming platforms such as Spotify, Apple Music, and YouTube Music rely heavily on

accurate genre tagging to power search tools, curated playlists, personalized recommendations, and algorithmic radio stations. Improved genre classification enables a more intuitive user experience and enriches music discovery pathways. Beyond the commercial sector, digital music libraries, cultural heritage archives, and academic music collections also stand to benefit from scalable tagging solutions that minimize the need for time-consuming manual annotation. For independent artists and content developers, this framework offers an accessible tool for automatically enriching metadata and enhancing content visibility in competitive digital marketplaces. Ultimately, this project contributes both to the development of novel computational tools and to the democratization of music classification technology.

For this project, the GTZAN Genre Collection, a benchmark dataset widely used for genre classification research, was used. This dataset consists of 1,000 tracks, each lasting 30 seconds and spanning ten distinct musical genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. The dataset was retrieved from Kaggle and includes audio recordings in WAV format. Each file was originally sampled at a higher rate, but for consistency and model compatibility, all files were downsampled to a standard 22.05kHz rate and converted stereo tracks to mono. Preprocessing also involved transforming the waveform audio into Mel-spectrograms using the Torchaudio library. The Mel-spectrogram provides a visually interpretable format that maps pitch and timing on a perceptually motivated scale, capturing both frequency and temporal characteristics. A standard configuration of 128 Mel bands, a fast Fourier transform (FFT) window of 2048, and a hop length of 512 were used. The generated spectrograms were further scaled to decibel units, ensuring uniform intensity values across examples. During this process, a small number of corrupted files were encountered, such as "jazz.00054.wav," which were removed from the training set to maintain data integrity.

The methodological foundation for this project is anchored in a well-established body of research supporting the efficacy of convolutional neural networks in the domain of musical signal processing. Numerous studies have demonstrated the utility of CNNs in learning hierarchical representations of complex auditory signals, especially when these signals are converted into two-dimensional spectrogram formats. For instance, the work of Pons et al. (2018) provided compelling evidence that CNNs trained end-to-end on Mel-spectrograms could achieve classification accuracies exceeding 85% in music tagging tasks. This study underscored the potential of using visual representations of audio, where convolutional kernels adeptly extract temporal and frequency-based features inherent in music. Dieleman and Schrauwen (2015) proposed an alternative perspective by applying CNNs directly to raw audio waveforms, bypassing traditional feature extraction pipelines. Their findings revealed the critical role of architectural tuning and the incorporation of data augmentation techniques—such as pitch shifting and time stretching—to enhance the generalizability of models trained on limited datasets. This research further highlighted the need for flexible architectures that can adapt to the diverse statistical properties of musical content. Additional scholarly contributions, including those by Choi et al. (2016) and Thickstun et al. (2017), have broadened the scope of deep learning applications in music classification by introducing deeper network structures and

self-supervised learning strategies capable of capturing high-level abstractions directly from musical input. Collectively, these studies form the conceptual and technical backdrop against which this project's approach is situated.

Informed by this body of research, a custom convolutional neural network was implemented using PyTorch to serve as the backbone of the genre classification pipeline. The architecture was intentionally designed with three sequential convolutional blocks, each characterized by an increasing number of filters—32, 64, and 128, respectively—to allow the network to progressively learn more abstract and complex representations of the input spectrograms. Within each block, the convolutional operation is followed by batch normalization, which stabilizes and accelerates the training process by reducing internal covariate shift. This is immediately succeeded by a rectified linear unit (ReLU) activation function, introducing non-linearity and enabling the model to capture intricate audio patterns. Max pooling layers are employed at the end of each block to downsample the feature maps and reduce computational overhead, while retaining the most salient features. To account for variability in input sizes and ensure a consistent output tensor shape prior to the dense layers, an adaptive average pooling layer is applied. This component standardizes the spatial dimensions of the feature maps, thereby facilitating a smooth transition to the fully connected portion of the network. The architecture reflects a deliberate balance between depth and regularization, structured to avoid overfitting while maintaining the capacity to learn from high-dimensional input data.

The classification head of the network is composed of three fully connected (dense) layers designed to perform high-level reasoning based on the learned convolutional features. The first dense layer projects the flattened feature tensor to a 256-dimensional representation, followed by a second layer that compresses this into 128 dimensions. The final layer is responsible for producing the genre classification output, consisting of ten logits corresponding to the ten musical genres in the GTZAN dataset. To mitigate overfitting and encourage generalization, dropout regularization is applied after both of the intermediate dense layers, with a dropout rate of 0.5. Batch normalization is also retained within the dense layers to further stabilize gradient updates during training. Model optimization is conducted using the Adam optimizer, known for its adaptive learning rate properties and strong empirical performance in deep learning tasks. The loss function selected for training is categorical cross-entropy, appropriate for multi-class classification problems with mutually exclusive labels. To prevent premature convergence and improve long-term learning, a ReduceLROnPlateau learning rate scheduler is incorporated, which reduces the learning rate upon stagnation in validation loss, using a patience parameter of three epochs. The model is trained over the course of 50 epochs using a batch size of 32, with performance metrics evaluated at each epoch to monitor progress and adjust hyperparameters dynamically.

To ensure an equitable distribution of genre classes across each experimental phase, the dataset was partitioned using a stratified split strategy: 80% of the data was allocated for training, 10% for validation, and the remaining 10% for testing. This partitioning method preserved class

proportions and minimized the risk of imbalance-induced bias in model evaluation. During training, both training and validation metrics—including loss and classification accuracy—were recorded and visualized to assess model convergence. The network exhibited strong learning behavior, with the training accuracy consistently improving and eventually surpassing 95%. Validation accuracy plateaued at approximately 82.5%, indicating successful generalization on unseen data without significant overfitting. Final testing of the best-performing model yielded a classification accuracy of 77.5%, outperforming several baseline implementations documented in the literature and demonstrating the practical viability of the chosen architecture and preprocessing strategy. A granular evaluation of model performance on a per-class basis revealed distinct patterns of strength and vulnerability in the classifier's predictive capability. The classical genre achieved the highest classification accuracy, with a success rate of 90%, likely attributable to its unique acoustic profile, characterized by consistent instrumentation, broad dynamic ranges, and minimal rhythmic variation. Similarly, the metal genre was identified with high confidence (88% accuracy), a result that may stem from its distinct high-frequency spectral features and percussive intensity. Conversely, the classifier exhibited greater difficulty distinguishing genres such as rock (60% accuracy) and blues (67% accuracy), which often share overlapping harmonic and rhythmic features with genres like country and jazz. These results point to a fundamental challenge in genre classification: the fuzzy boundaries between musical categories, which are frequently shaped by cultural, historical, and production-related factors. A confusion matrix provided additional insight into these misclassifications, revealing clusters of inter-genre ambiguity, particularly among popular music genres with shared stylistic elements. These findings underscore the limitations of a purely spectral approach and suggest that genre classification may benefit from the incorporation of additional contextual or temporal cues.

The model's capacity to generalize was further substantiated by secondary performance metrics such as macro-averaged precision, recall, and F1 scores. These metrics, particularly valuable in multi-class settings with class imbalance, reinforced the model's relatively balanced performance across most categories. Nonetheless, several limitations emerged that warrant critical reflection. Chief among these is the constrained size of the GTZAN dataset, which, while useful for benchmarking, lacks sufficient diversity in terms of artist representation and recording conditions. This limitation restricts the breadth of learned features and hampers generalizability to real-world, heterogeneous audio corpora. Furthermore, the current approach, grounded in convolutional operations on spectrogram inputs, captures spatial dependencies within short temporal windows but lacks mechanisms for modeling long-range temporal dynamics. As a result, the classifier is not equipped to recognize higher-order temporal patterns or structural elements that span across multiple seconds of audio, which are often crucial for genre identification. The absence of temporal modeling architectures such as recurrent neural networks or attention mechanisms likely imposed an upper bound on achievable performance, especially for genres characterized by temporal evolution rather than isolated timbral features.

Despite these limitations, the findings of this project underscore the technical viability and practical effectiveness of convolutional neural networks as a foundational approach for

automated music genre classification. The full pipeline—from initial audio preprocessing and spectrogram generation to model training, evaluation, and prediction—exhibited a high degree of stability, transparency, and flexibility. This reliability is particularly significant given the inherent challenges of working with small and imbalanced datasets, and it affirms the utility of CNN-based architectures in extracting musically relevant features from spectrogram representations. The interpretability of the system further reinforces its value, as architectural design choices such as layered convolutional blocks and adaptive pooling facilitate insight into the hierarchical feature extraction process. Regularization techniques, including dropout and batch normalization, proved essential in maintaining generalization performance, particularly in mitigating overfitting across a limited number of training examples. These insights highlight the importance of intentional design in model architecture and preprocessing pipelines.

To further advance this work, future efforts could explore the integration of temporal modeling techniques, such as recurrent neural networks (RNNs), long short-term memory (LSTM) units, or self-attention mechanisms used in transformer architectures. These models are well-suited to capturing long-range dependencies in audio and could offer meaningful performance gains by learning temporal transitions that span across musical bars or phrases. Additionally, expanding the training data through audio data augmentation—such as pitch shifting, time stretching, background noise injection, or spectral masking—could enhance robustness and improve generalization to real-world, noisy, or non-canonical audio inputs. This classification framework may also be extended to other domains within music information retrieval. Beyond genre identification, the underlying methodology could be adapted for tasks such as instrument recognition, mood or sentiment classification, artist identification, or even structural segmentation of musical works. By refining and generalizing this approach, it may serve as a modular and extensible tool within the broader MIR ecosystem, contributing to both scholarly research and industry applications aimed at enhancing the organization, discovery, and accessibility of digital music content.

References

Choi, Keunwoo, György Fazekas, and Mark Sandler. 2016. “Automatic Tagging Using Deep Convolutional Neural Networks.” arXiv preprint arXiv:1606.00298. https://arxiv.org/abs/1606.00298.

Dieleman, Sander, and Benjamin Schrauwen. 2015. “End-to-End Learning for Music Audio.” arXiv preprint arXiv:1506.03194. https://arxiv.org/abs/1506.03194.

Librosa Development Team. 2023. Librosa: Audio and Music Processing in Python. https://librosa.org/.

Paszke, Adam, Sam Gross, Francisco Massa, et al. 2019. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” Advances in Neural Information Processing Systems 32: 8026–8037.

Pons, Jordi, Xavier Serra, and Joan Serrà. 2018. “End-to-End Learning for Music Audio Tagging at Scale.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (8): 1680–1693. https://doi.org/10.1109/TASLP.2018.2836343.

Thickstun, John, Zaid Harchaoui, and Sham Kakade. 2017. “Learning Features of Music from Scratch.” International Conference on Learning Representations. https://openreview.net/forum?id=ryFm1j-C-.

TorchAudio Contributors. 2022. TorchAudio: Audio Data Loading and Augmentation for PyTorch. https://pytorch.org/audio/stable/index.html.

Tzanetakis, George, and Perry Cook. 2002. “GTZAN Genre Collection.” Accessed via Kaggle. https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification.