FDS: Genre music identification 2025

Author One
Giannicola Mongelli

Author Two
Francesco Nardiello

Author Three
Karim Ciacciarelli

1 Dataset and Pre-Processing

The project utilizes the GTZAN dataset, consisting of 1,000 audio tracks with a duration of 30 seconds each, classified into ten musical genres. During the initial inspection phase, a corrupted file (jazz00054.wav) was identified and removed. Following this cleaning operation, the dataset was partitioned into a training set (80%) and a test set (20%).
An initial analysis revealed that the quantity of original data was insufficient to train deep learning models with satisfactory performance. To overcome this issue and increase the number of available examples, a track segmentation strategy was adopted. In order to identify the optimal configuration, several experimental trials were conducted by subdividing the original tracks into segments of varying lengths (3, 5, and 10 seconds). This procedure significantly expanded the dataset and allowed for an evaluation of the impact of segment duration on classification accuracy, while keeping the nature of the audio signal unchanged. Furthermore, during the model training phase, the primary training set was further partitioned using an 80-20 split to create a dedicated validation set.

2 Spectrograms and Waveforms

Based on the audio segments, two distinct data representations were generated: spectrograms and waveforms, using Librosa library [1].
Spectrograms, obtained through the **Short-Time Fourier Transform (STFT)**, transform the audio signal into a two-dimensional matrix (time-frequency). The transformation is defined as [2]:

$$STFT\{x(t)\}(\tau,\omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-j\omega t}dt$$

Where:

- $x(t)$ represents the original audio signal in the continuous time domain.
- $w(t-\tau)$ is the *window function*; it slides along the signal, translated by a time factor τ. Its function is to isolate a short segment of the signal around the instant τ, suppressing the signal's contributions outside of this interval. This ensures the temporal resolution that is missing in the standard Fourier transform.
- $e^{-j\omega t}$ acts as a "probe" that correlates the segment isolated by the window with a sinusoid of angular frequencysoide di frequenza angolare ω.

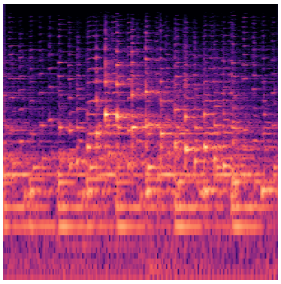
Following the calculation of the transform, the result is a matrix of complex numbers. To generate the final image (the spectrogram), the magnitude of the signal is calculated and a logarithmic scale transformation (Decibel) is applied. The formula used is:

$$S_{dB}(\tau,\omega) = 20 \cdot \log_{10}\left(\frac{|STFT(\tau,\omega)|}{\max_{\tau',\omega'}|STFT(\tau',\omega')|}\right)$$

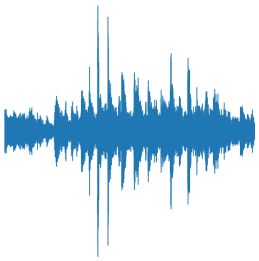
Waveforms, instead, maintain the signal in the time domain as one-dimensional vectors. The waveform preserves the raw amplitude data, capturing the dynamics, transients, and rhythmic patterns with high temporal resolution. These representations are derived by applying **Amplitude Normalization** to the previously segmented audio, where each segment is scaled by its maximum absolute value ($y_{norm} = \frac{y}{\max|y|} + \epsilon$, [3]) to ensure numerical stability. This allows the network to identify rhythmic "spikes" and melodic envelopes, effectively treating temporal dynamics as visual patterns that complement the frequency-domain information. Figure 1 illustrates a dual-domain representation of a 5-second Classical audio segment, contrasting its spectral and temporal characteristics.

3 Models

In order to identify the most effective strategy for audio spectrogram classification, experimentation was conducted on several Deep Learning architectures.

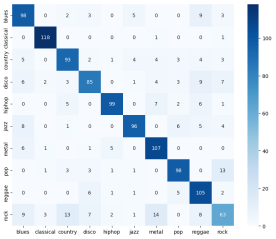


a) Spectrogram

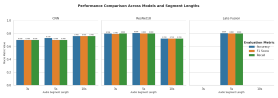


b) Waveform

Figure 1: Comparative Analysis of a) Spectrogram and b) Waveform Representations for a 5-Second Classical Audio Segment.



a) ResNet18 5sec confusion matrix



b) Models results

Figure 2: a) Confusion matrix of ResNet18 5sec, b) Results of each model across all segment lengths

3.1 CNN

This model is a custom 6-layers trainable architecture built from scratch, designed to process 128 $\times$128$\times$3 spectrogram inputs. The model utilizes two sequential convolutional blocks, each containing a 3$\times$3 convolution layer, Batch Normalization for training stability and ReLU activation. The first block extracts 16 filters, while the second increases the depth to 32. Then, each block is followed by a Max Pooling layer that halves the spatial dimensions, compressing the input down to a 32×32 feature map. After flattening the data, the network employs a Fully Connected layer with 256 neurons. A Dropout rate of 40% is applied to prevent overfitting before the final linear layer maps the features to the 10 musical genres.

3.2 ResNet18

The second strategy involves a Transfer Learning approach [4]. The model utilizes ResNet18 initialized with weights from the ImageNet dataset. This allows the network to leverage pre-existing visual feature extractors. The architecture is characterized by skip connections that bypass one or more layers, effectively mitigating the vanishing gradient problem and allowing for more stable training. To adapt the model to the GTZAN dataset, the original fully connected layer was replaced with a custom sequential head. This includes a Linear layer with 256 neurons, a ReLU activation and a Dropout layer to enhance generalization and reduce overfitting. The model is trained using the Adam optimizer with a learning rate and CrossEntropyLoss. Input spectrograms are resized to 128$\times$128 and normalized using ImageNet statistics to ensure compatibility with the pre-trained weights.

3.3 Late Fusion Multi-Modal [5]

The final strategy employs an architecture which integrates both spectral and temporal data to improve genre classification. The process is structure into three main phases:

- `Dual-Stream Feature Extraction`: The model utilizes two parallel ResNet18 acting as feature extractors. Both are initialized with pre-trained weights, but their final fully connected layers are replaced with an Identity layer;
- `Late Fusion Mechanism`: The 512-D feature vectors from both streams are concatenated into a single 1024-D dimensional representation. This combined vector is then fed into a final classification head, consisting of a linear layer, a ReLU activation and Dropout (0.5) to prevent overfitting before the final output;
- `Multi-Modal Data Management`: A custom class ensures synchronized data loading by fetching the corresponding spectrogram and waveform image for each audio segment. Furthermore, the system is optimized using Adam optimizer with a reduced learning rate.

4 Results

The comparative analysis identified the ResNet18 architecture trained on spectrograms with 5-second segmentation as the most performing model among the tested configurations, as shown in Figure 2. Temporal segmentation provided the ideal data density, whereas the multi-modal approach suffered from the introduction of noise resulting from the conversion of waveforms into images.
The model's efficacy is visually confirmed by the Confusion Matrix: the network is nearly infallible on timbrally distinct genres such as Classical and Metal. Conversely, Rock emerges as the primary challenge: its hybrid stylistic nature leads to frequent misclassifications as Metal, Country, or Pop. This highlights how semantic overlap between genres remains the most significant hurdle for automated classification.

References

[1] Librosa https://librosa.org/doc/latest/index.html

[2] Short-time Fourier transform https://en.wikipedia.org/wiki/Short-time_Fourier_transform

[3] Peak Normalization https://hackaudio.com/tutorial-courses/learn-audio-programming-table-of-contents/digital-signal-processing/amplitude/peak-normalization/

[4] ResNet18 https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html

[5] Github repo https://www.cs.cmu.edu/~morency/MMML-Tutorial-ACL2017.pdf

[6] Github repo https://github.com/francesco2706/Auto_identification