SAPIENZA
UNIVERSITÀ DI ROMA

MUSIC GENRE REVEAL

A CLASSIFICATION MODEL FOR AUDIOS
FDS 25/26

NARDIELLO FRANCESCO
CIACCIARELLI KARIM
MONGELLI GIANNICOLA

What?

A classification model capable of recognizing music genres.

Why ?

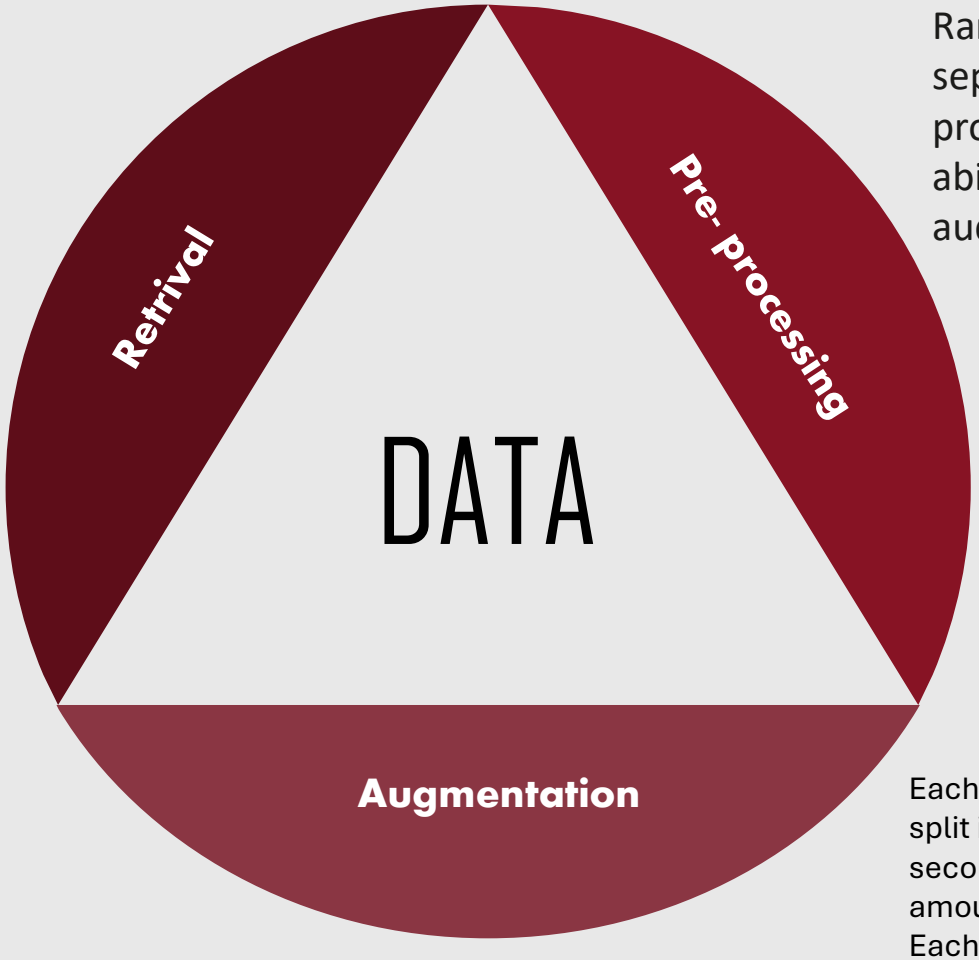
Accurate genre classification enables users to organize massive music libraries efficiently. It is also used as a foundation for recommendation on streaming platforms

How ?

By feeding into the model audio files converted into spectrograms

DATA

The dataset used is available on kaggle at **GTZAN Dataset.** It is collection of 10 genres with 100 audio files each, all having a length of 30 seconds



Random split of the dataset into separate training and test sets to properly evaluate the model's ability to generalize to unseen audio samples

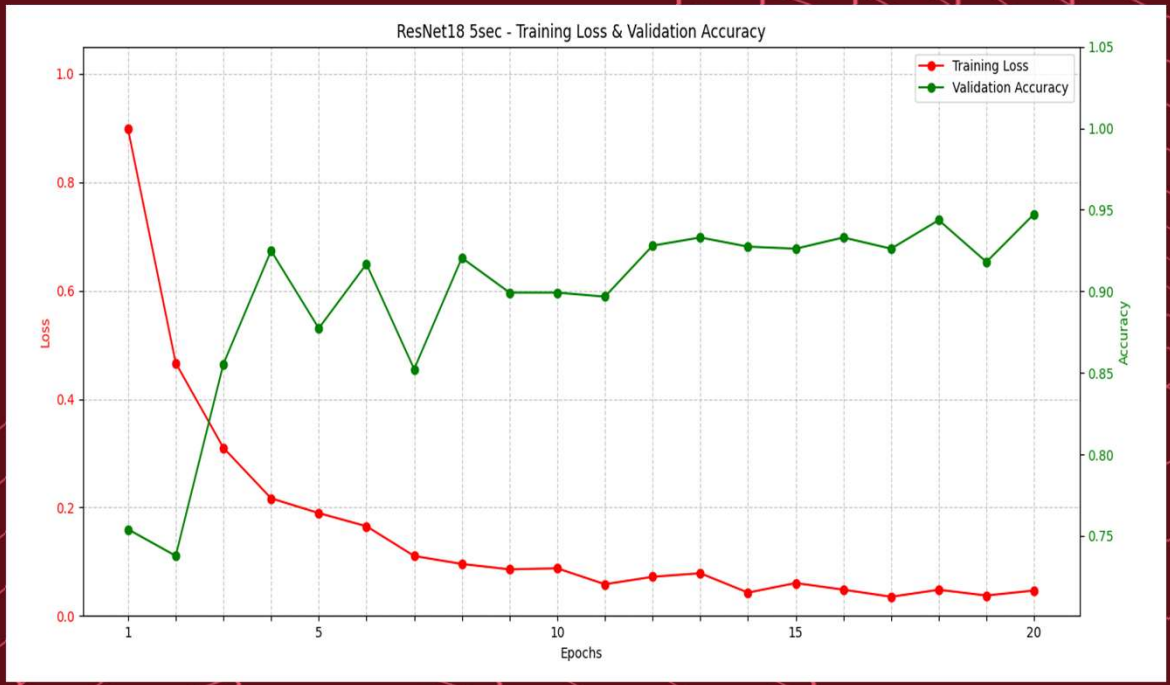
Each lengthy audio track of 30 seconds was split into multiple shorter segments (3, 5, 10 seconds), in order to increase the total amount of data, creating separate datasets. Each resulting segment was then converted into a **spectrogram**, transforming the raw audio into a time-frequency image suitable for the neural network

Models (1/2)

CNN and ResNet18 separately

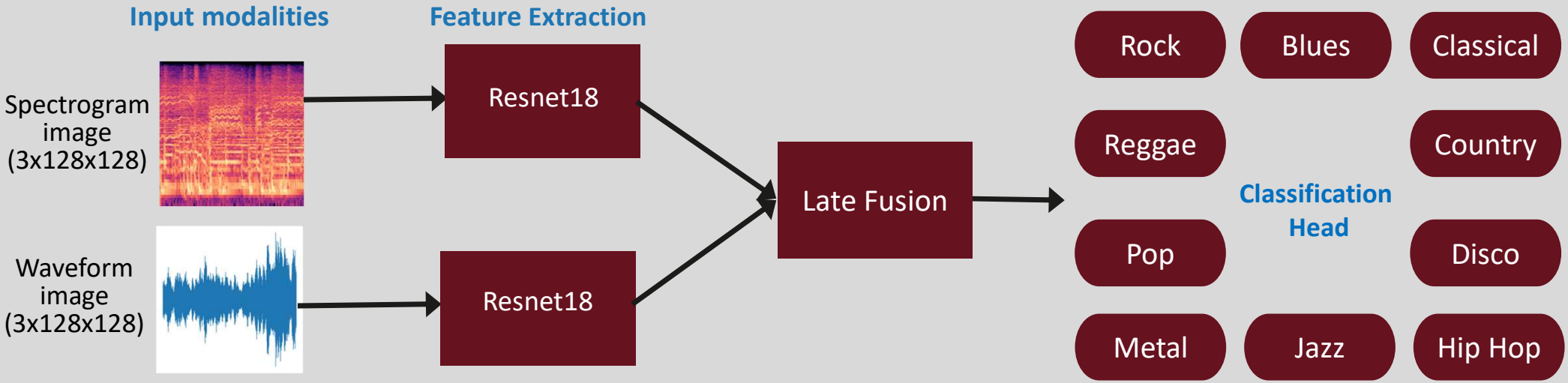
- Image size: 128
- Layers: 6 (CNN), 18 (ResNet)
- Epochs: 20

ResNet18 yielded the best results in terms of evaluation metrics, leveraging the vast visual knowledge acquired on ImageNet to ensure optimal generalization.

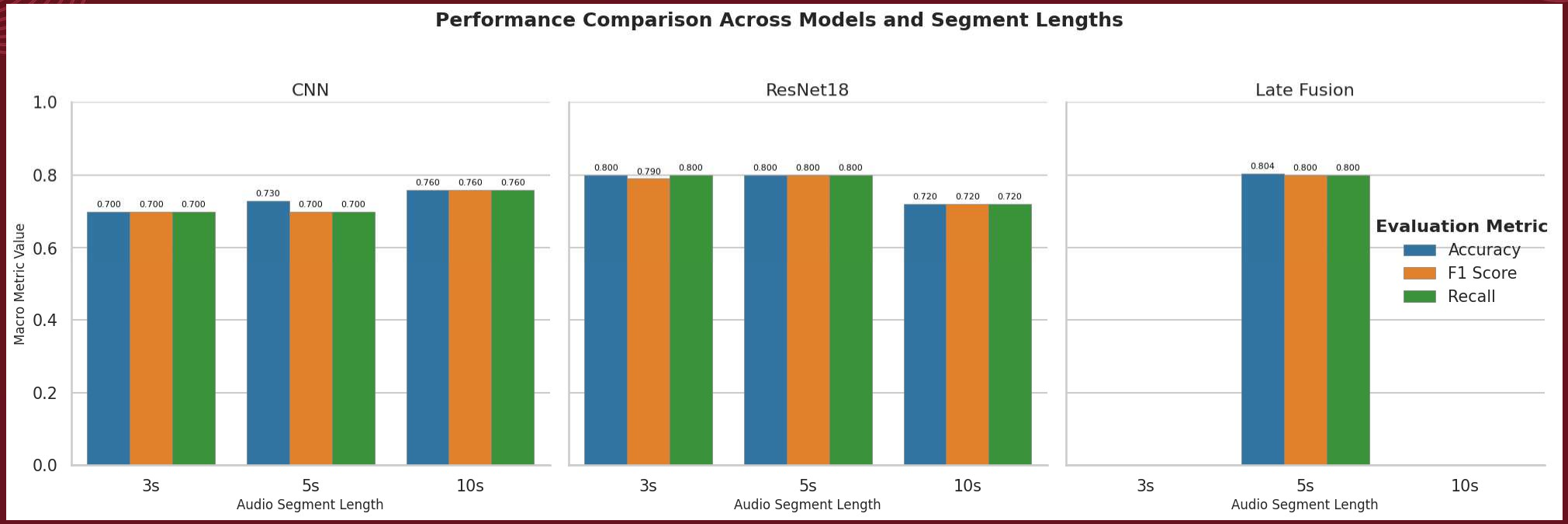


Models (2/2) - Late Fusion Multi Modal

- **Late fusion**: approach where individual models are trained on separate modalities, and their predictions are combined at a later stage.
- **Multi modal** = more input.
- **Implementation**: two parallel ResNet18, each acting as a dedicated features extractor on spectrograms and waveforms separately.



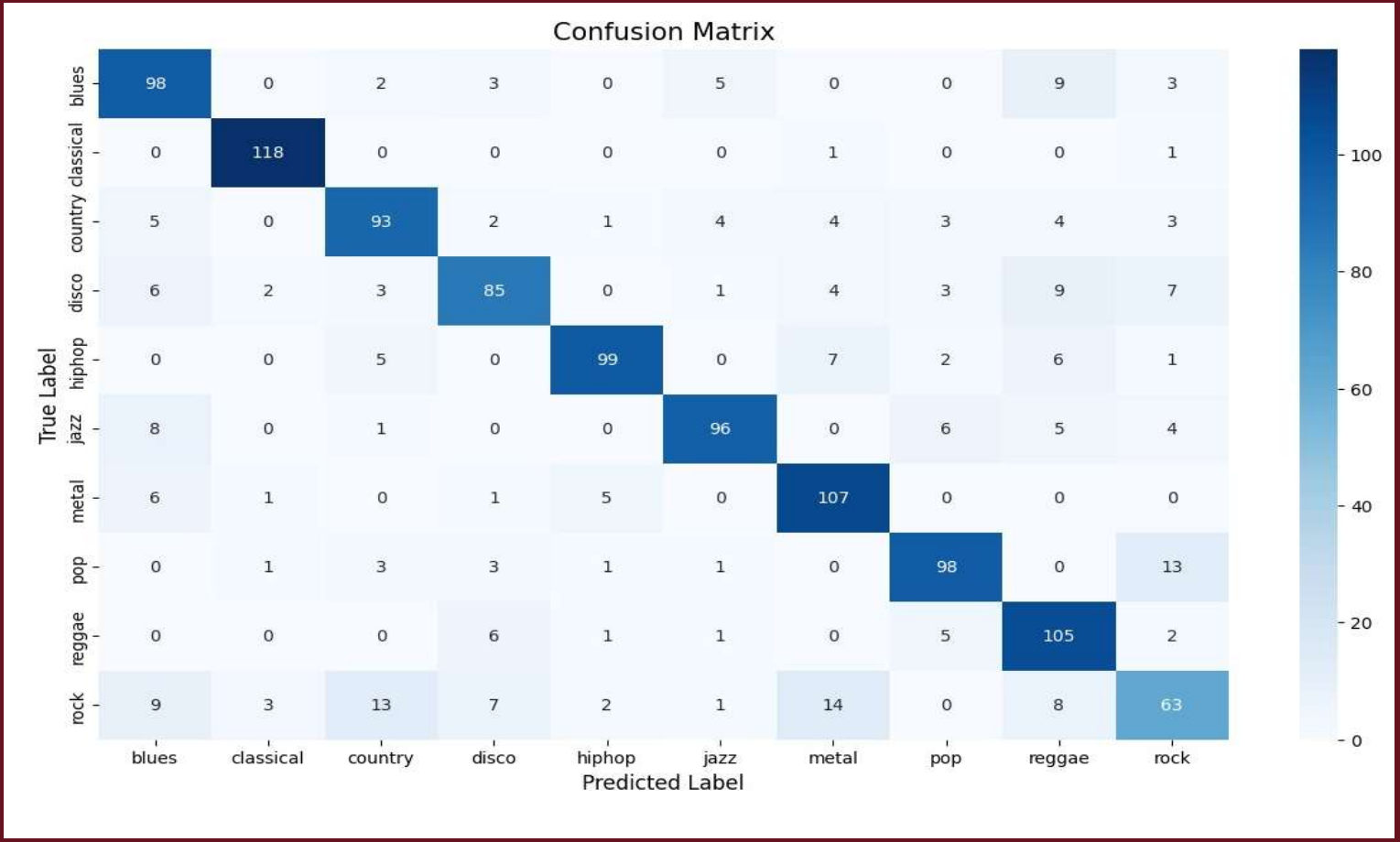
Results



ResNet18 was the strongest baseline, achieving 0.8 Accuracy at 3s and 5s, validating the efficacy of Transfer Learning. However, ResNet18 exhibited an unexpected performance drop on the longest segments (10s).

Late Fusion showed no significant accuracy difference versus the ResNet18 baseline, confirming that the simple concatenation strategy failed to extract meaningful synergy from the Waveform modality.

Quick focus on Resnet18



Classical and metal results in the highest scores which are the «extremis» of the dataset. Rock instead is one of the weakest often mistaken for blues or country probably due to similiraty between genres.

Conclusions and Future Works

Conclusions:

The Fusion failed because the Waveform Image introduced noise and redundancy. Conversely, the SpectroGram captures complex frequency and timbral features.

Future Works:

- Aim to achieve better performance in fusion modality.
- Test the model's performance on a large, more diverse dataset to ensure its readiness for real-world deployment and scalability.

References

Music genre classification with parallel convolutional neural networks
https://www.nature.com/articles/s41598-025-90619-7

Music Genre Classification Using Convolutional Neural Networks
https://github.com/crlandsc/Music-Genre-Classification-Using-Convolutional-Neural-Networks

Kaggle Dataset
https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

Image by Daniel Schludi on unplash

Contacts

Project page
https://github.com/francesco2706/Auto_identification