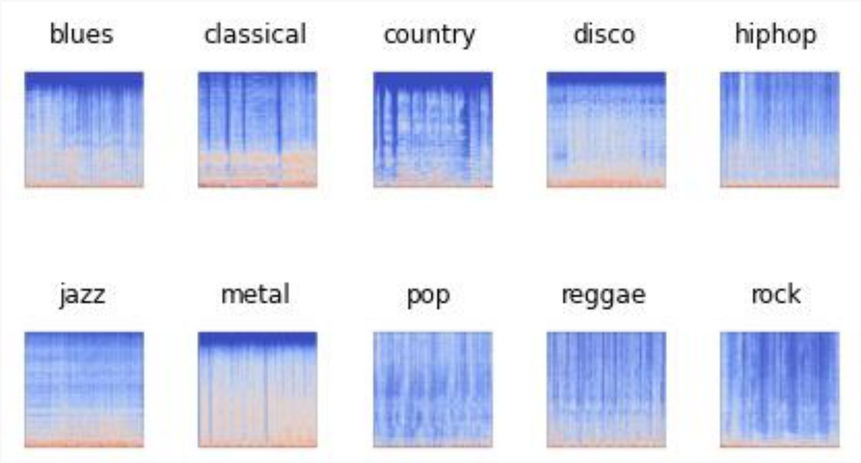
Music Genre Classification

Xarxes Neuronals i Aprenentatge Profund

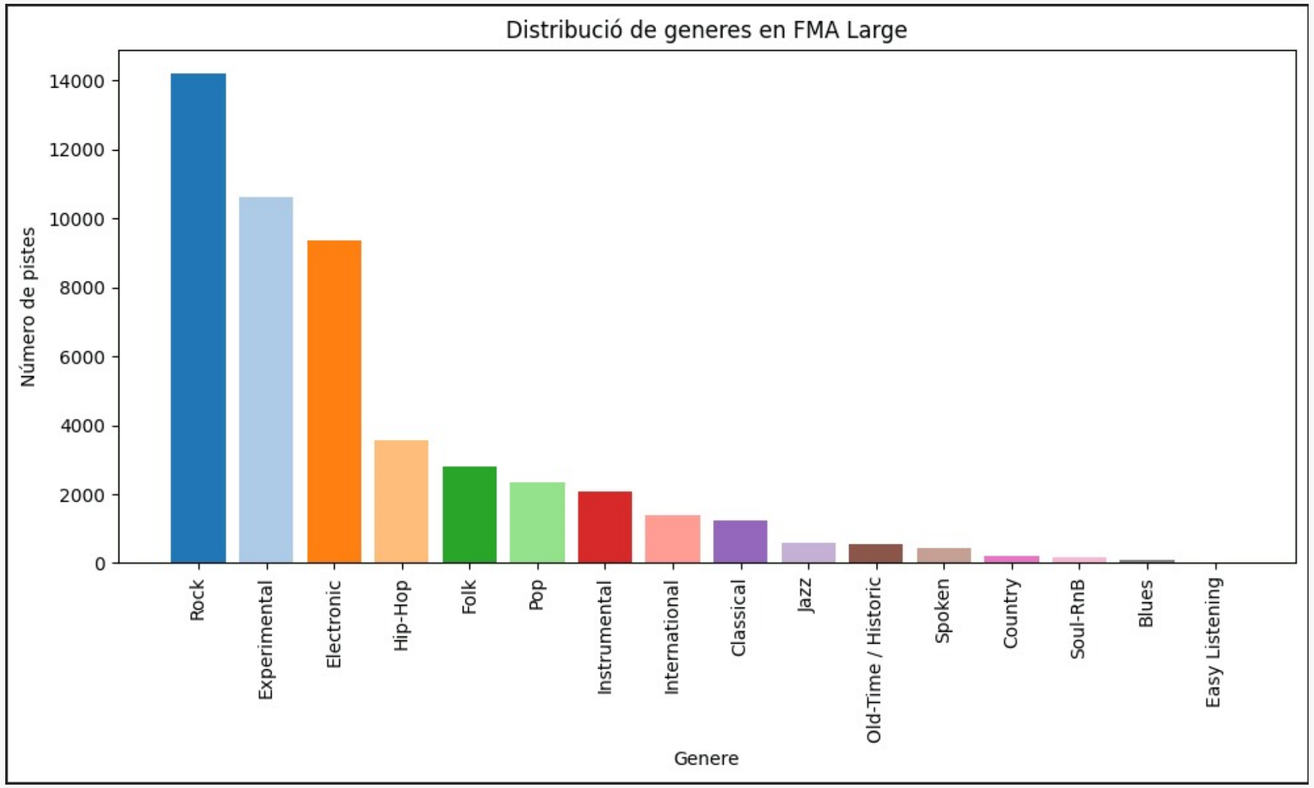
Nerea de la Torre, Mara Montero, Júlia Morán i Adrián Prego

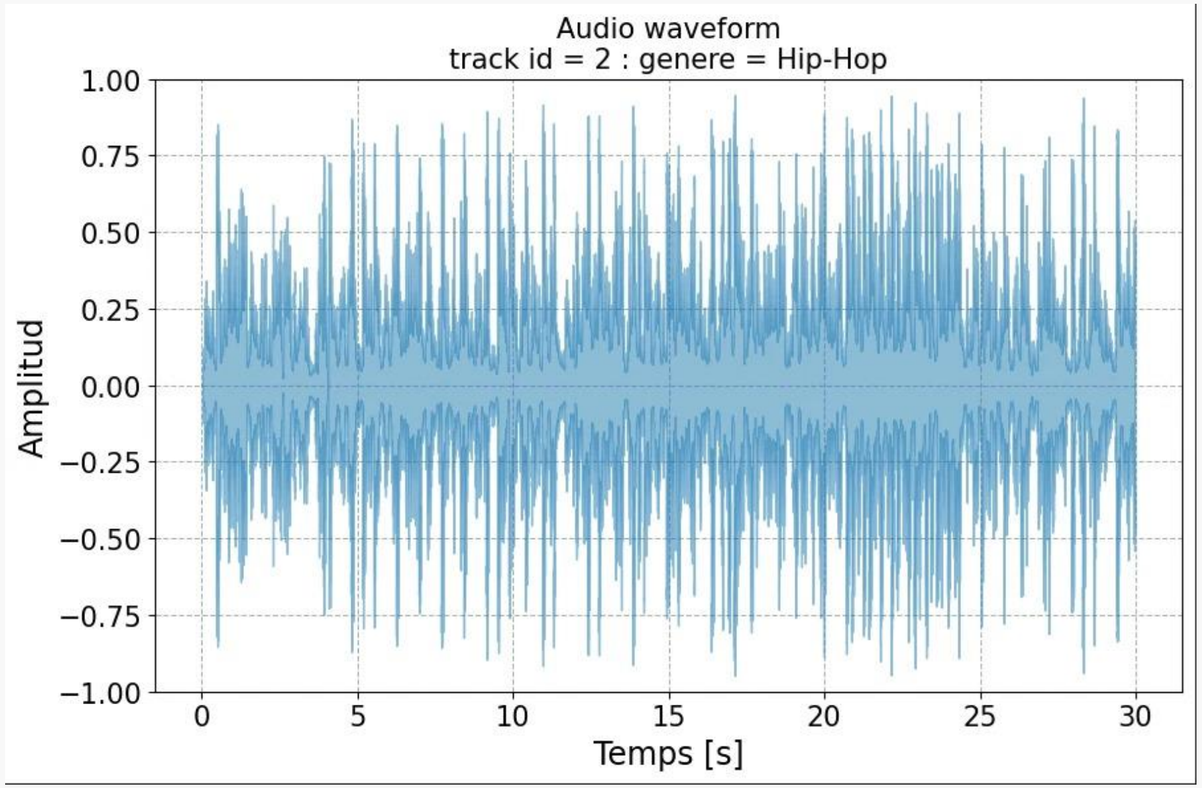
Introducció i objectiu

Comparar rendiment de diferents arquitectures per a la classificació de gèneres musicals

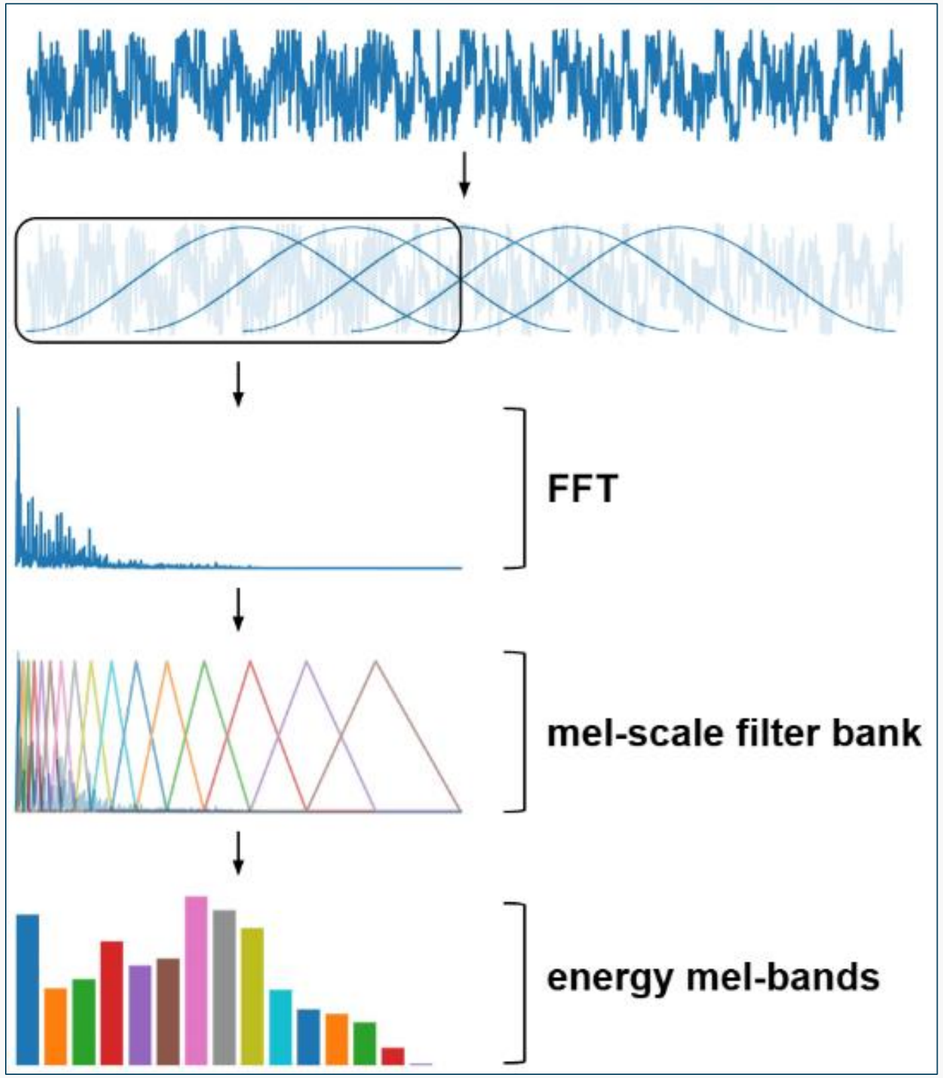


Dataset i EDA

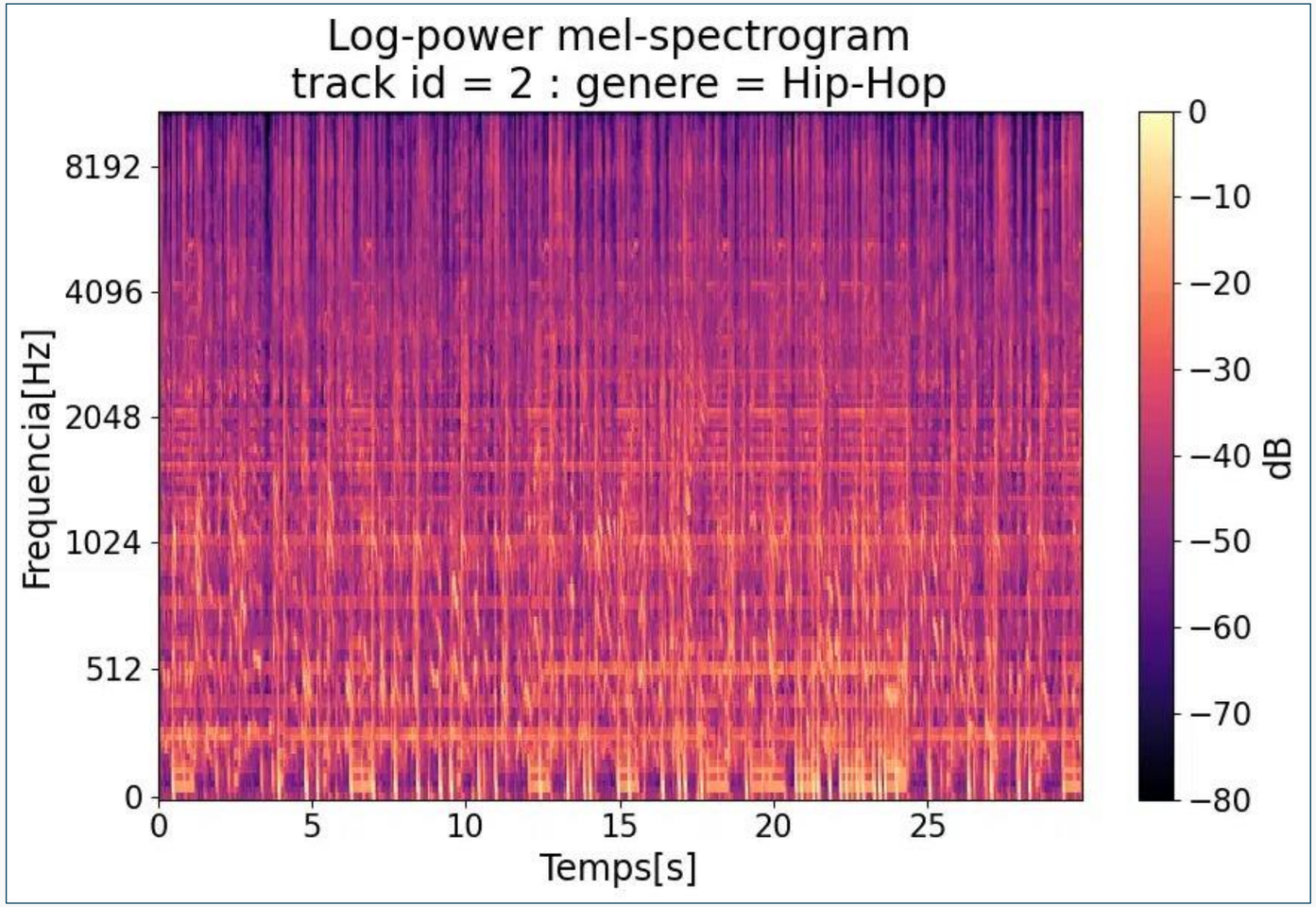




Feature Extraction

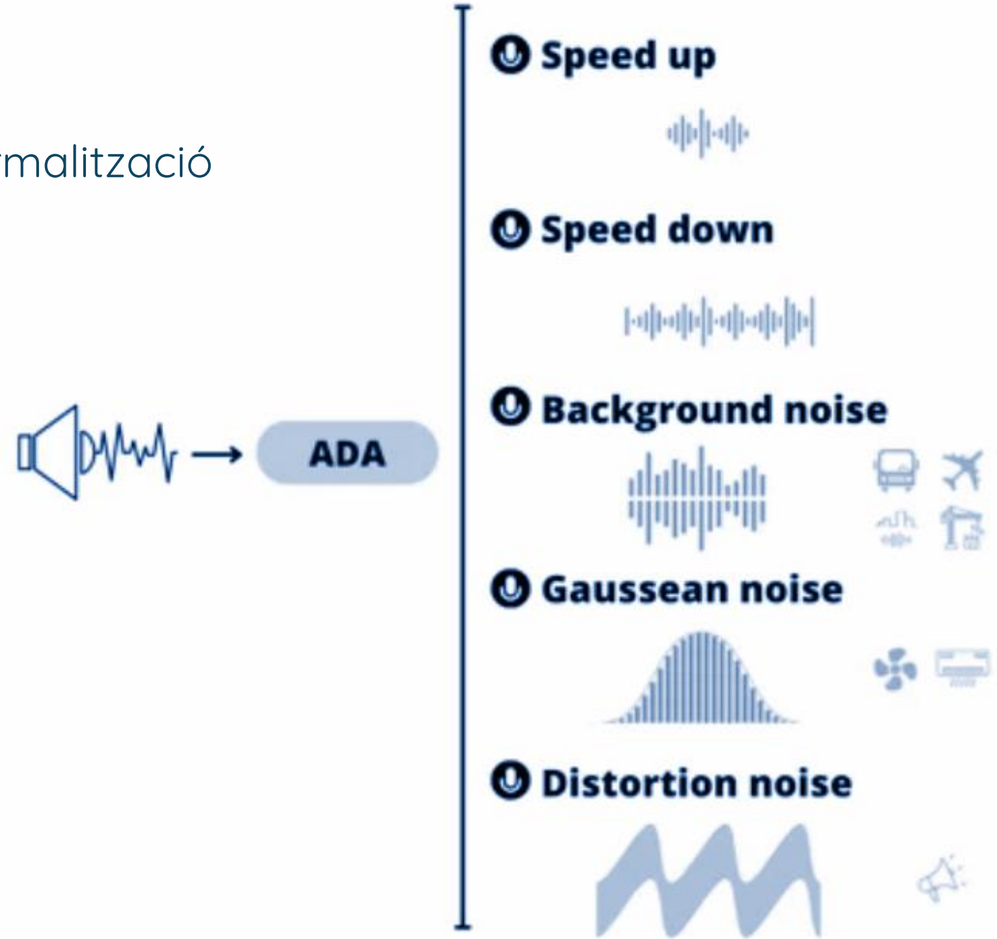


Àudio .mp3 -> espectrograma mel



Preprocessing

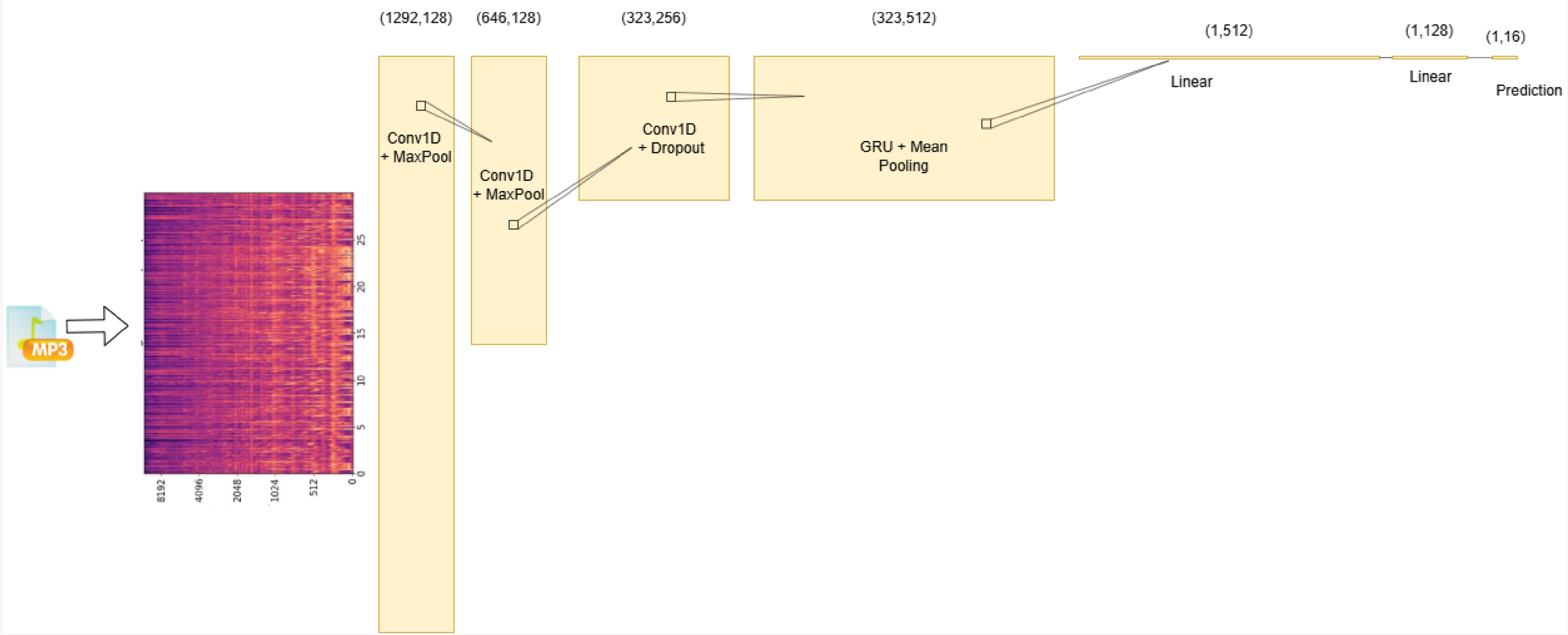
- Transformació a dB + normalització
- Eliminació NaN
- Data augmentation



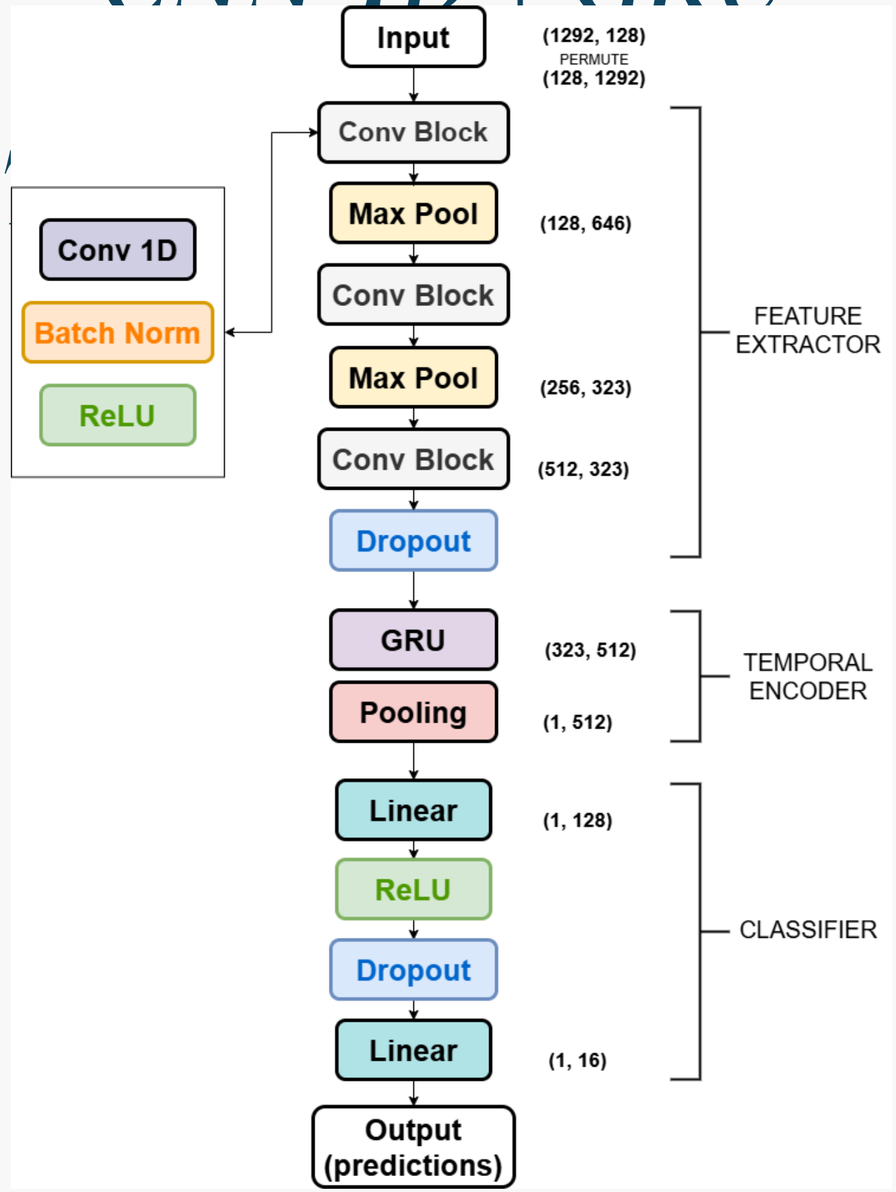
Arquitectures

- CNN(**1D**) + RNN **GRU**
- CNN(**1D**) + RNN **LSTM**
- CNN(**2D**) + RNN **LSTM**
- CNN(**2D**) + RNN **GRU**

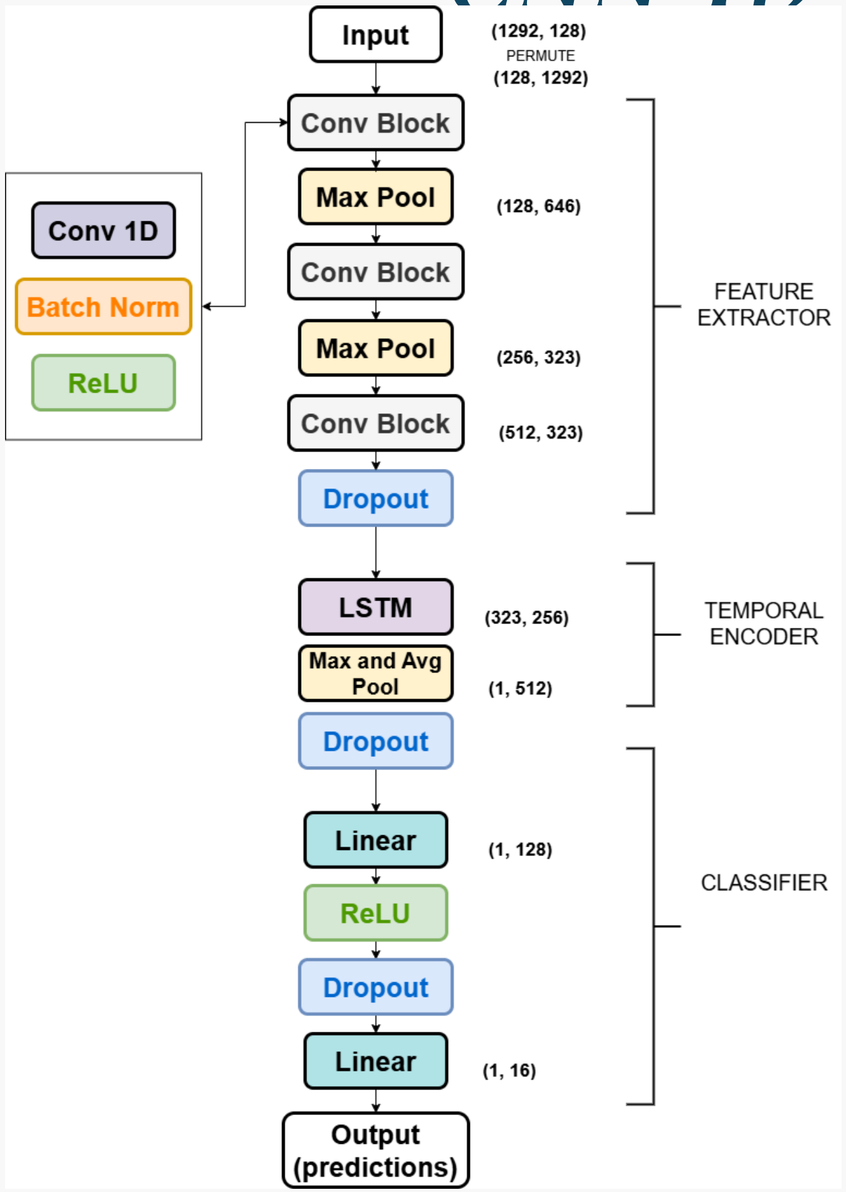
CNN-1D


(1292,128)
(646,128)
(323,256)
(323,512)
(1,512)
(1,128)
(1,16)
Conv1D
+ MaxPool
Conv1D
+ MaxPool
Conv1D
+ Dropout
GRU + Mean
Pooling
Linear
Linear
Prediction
MP3
8192
4096
2048
1024
512
0
5
10
15
20
25

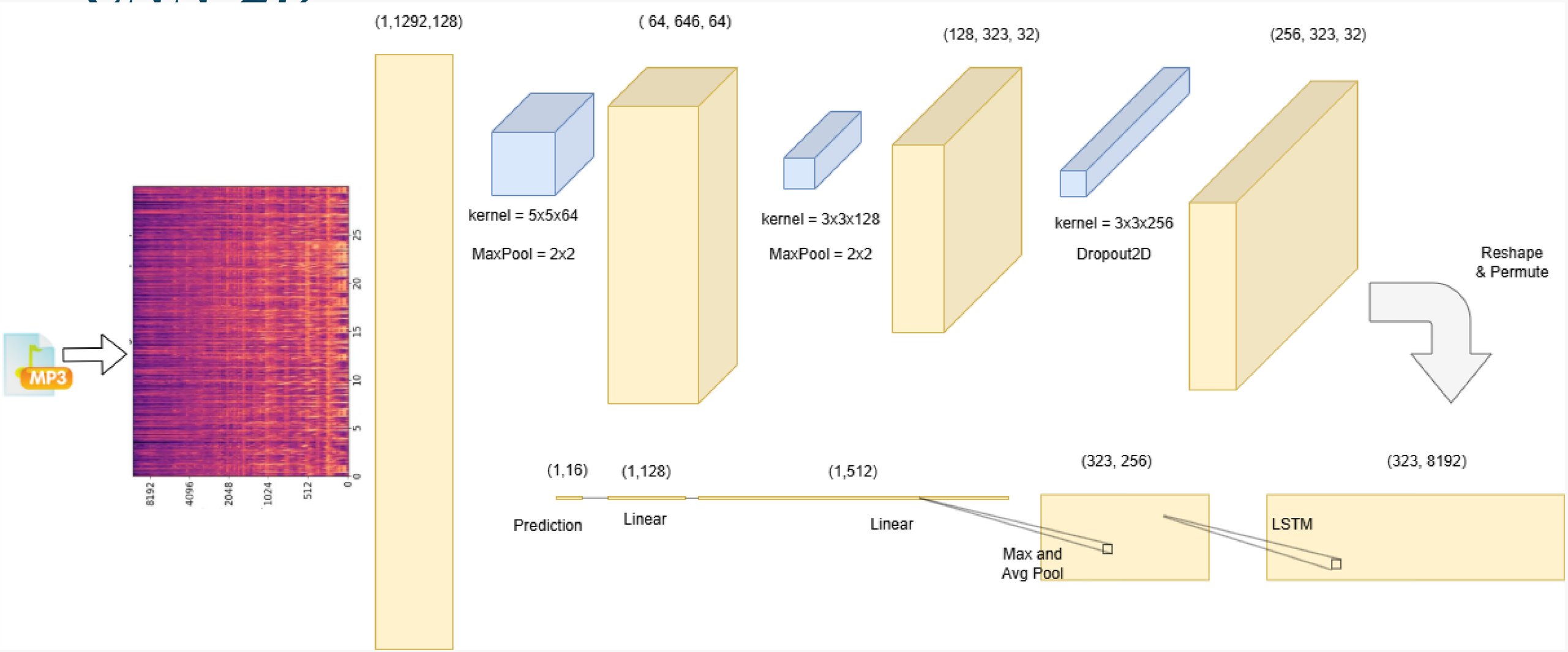
CNN-1D + GRU



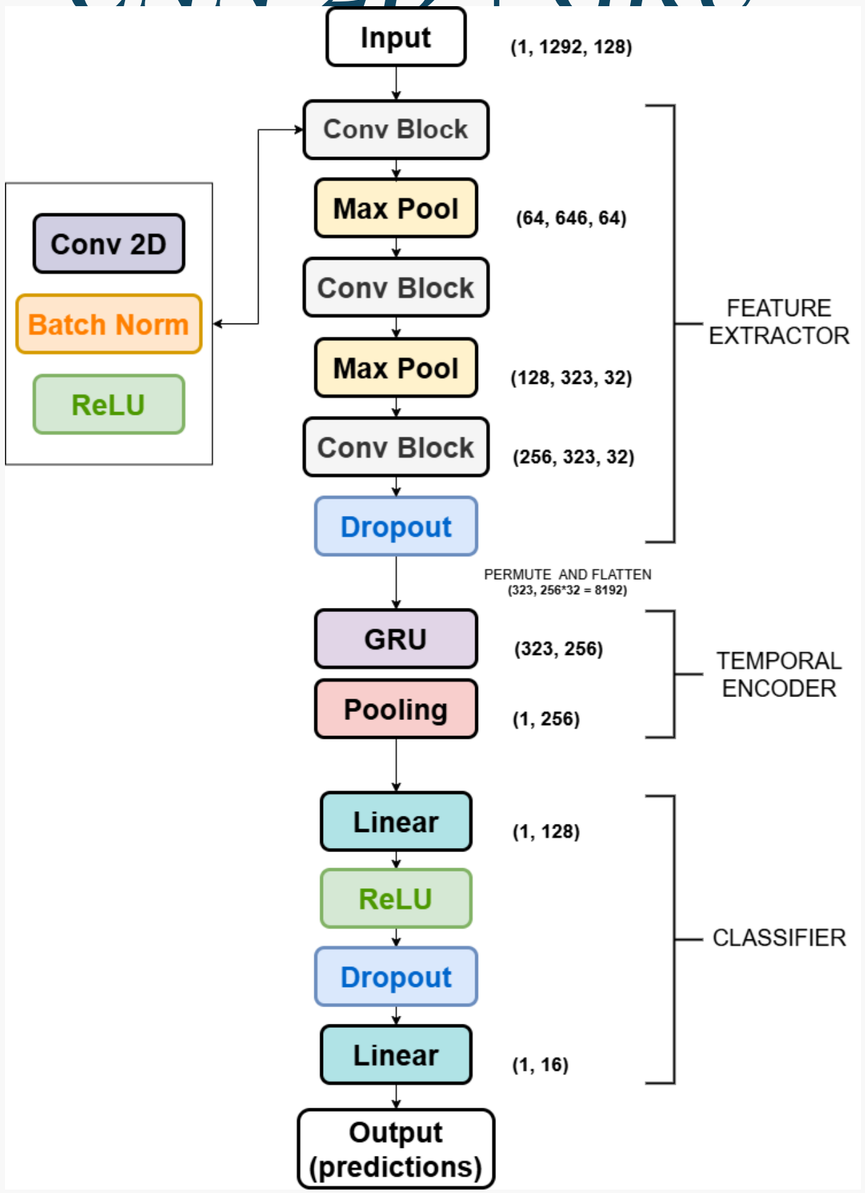
CNN-1D +



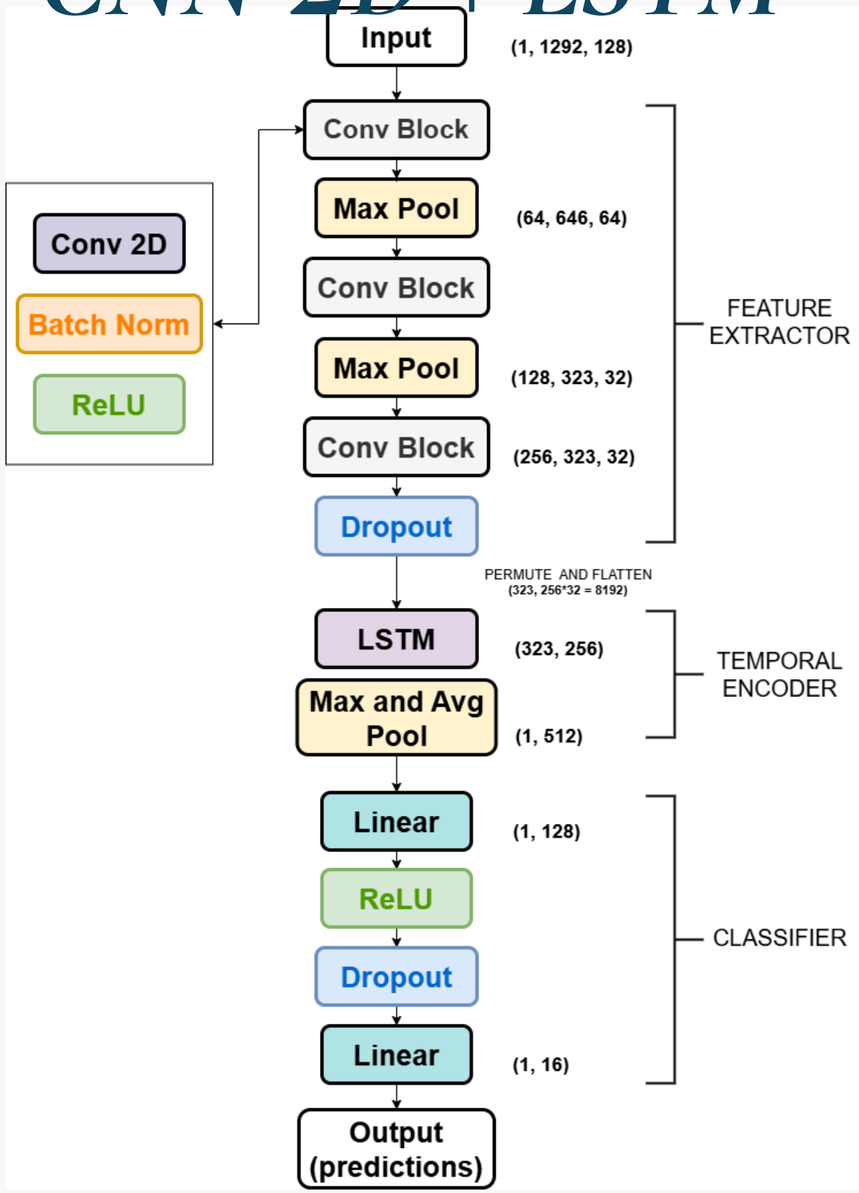
CNN-2D


(1,1292,128)
(64, 646, 64)
(128, 323, 32)
(256, 323, 32)
kernel = 5x5x64
MaxPool = 2x2
kernel = 3x3x128
MaxPool = 2x2
kernel = 3x3x256
Dropout2D
Reshape
& Permute
MP3
8192
4096
2048
1024
512
0
5
10
15
20
25
(1,16)
(1,128)
(1,512)
(323, 256)
(323, 8192)
Prediction
Linear
Linear
Max and
Avg Pool
LSTM

CNN-2D + GRU



CNN-2D + LSTM



Training

Split: Train 80%, Validation 10%, Test 10%

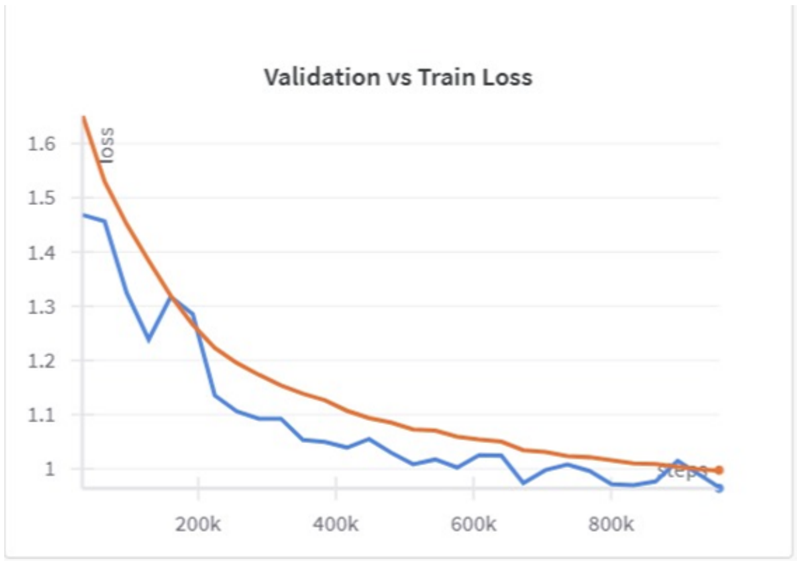
- Mètriques: Accuracy, F1 macro i weighted
- Multi-class **Cross Entropy** Loss
- **Adam** Optimizer + weight decay
- Learning rate adaptat a cada model
- **Bayesian** Search

$$\text{F1-macro} = \frac{1}{C}\sum_{c=1}^{C} 2 \times \frac{\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} \text{TP}_c + \text{TN}_c}{\sum_{c=1}^{C} (\text{TP}_c + \text{TN}_c + \text{FP}_c + \text{FN}_c)} \times 100\%$$

$$\text{F1-weighted} = \sum_{c=1}^{C} \frac{n_c}{\sum_{k=1}^{C} n_k} \cdot \left(2 \times \frac{\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \right)$$

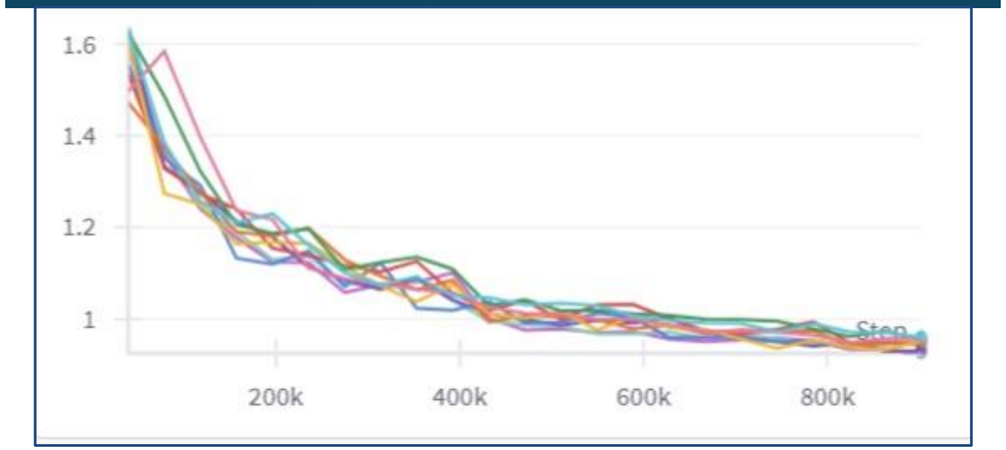
Primers models i millores



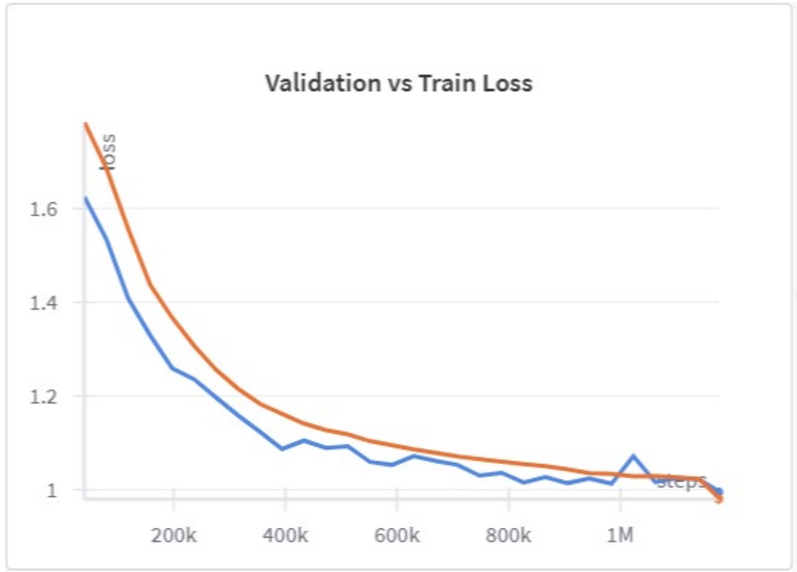
CNN(1D)-GRU

- weight decay
- ReduceLROnPlateau
- Early stopping

```yaml
# sweep.yaml
parameters:
  epochs:
    value: 30

  batch_size:
    values: [16, 32, 64, 128]

  learning_rate:
    values: [1e-4, 3e-4, 5e-4, 7e-4, 1e-3]

  hidden_dim:
    values: [64, 128, 256]

  dropout:
    values: [0.2, 0.3, 0.4, 0.5]

  num_layers: #en el RNN

    values: [1, 2]
```

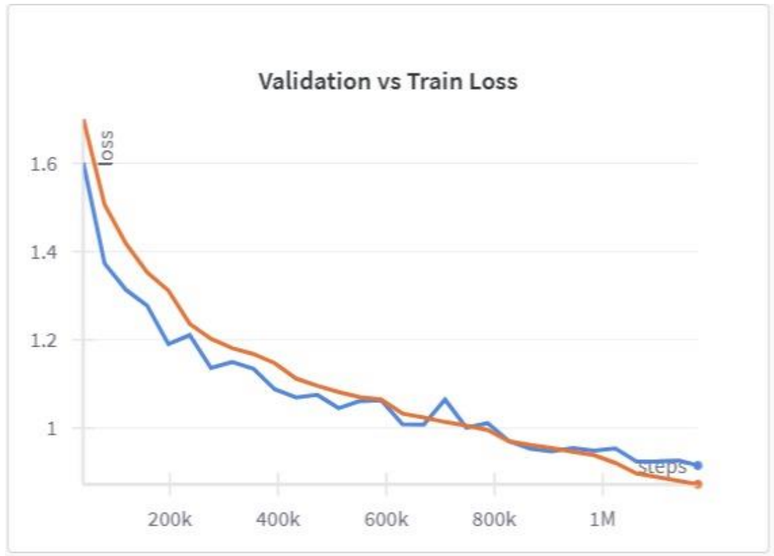


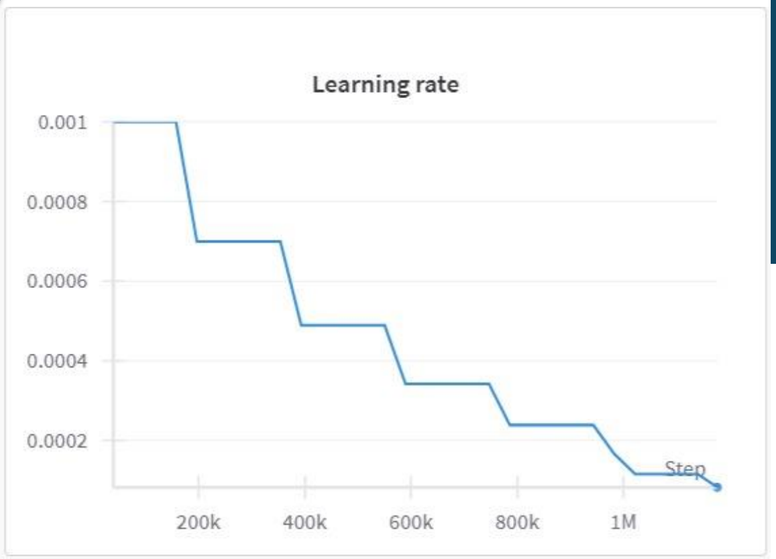
Primers models i millores



CNN(1D)-LSTM

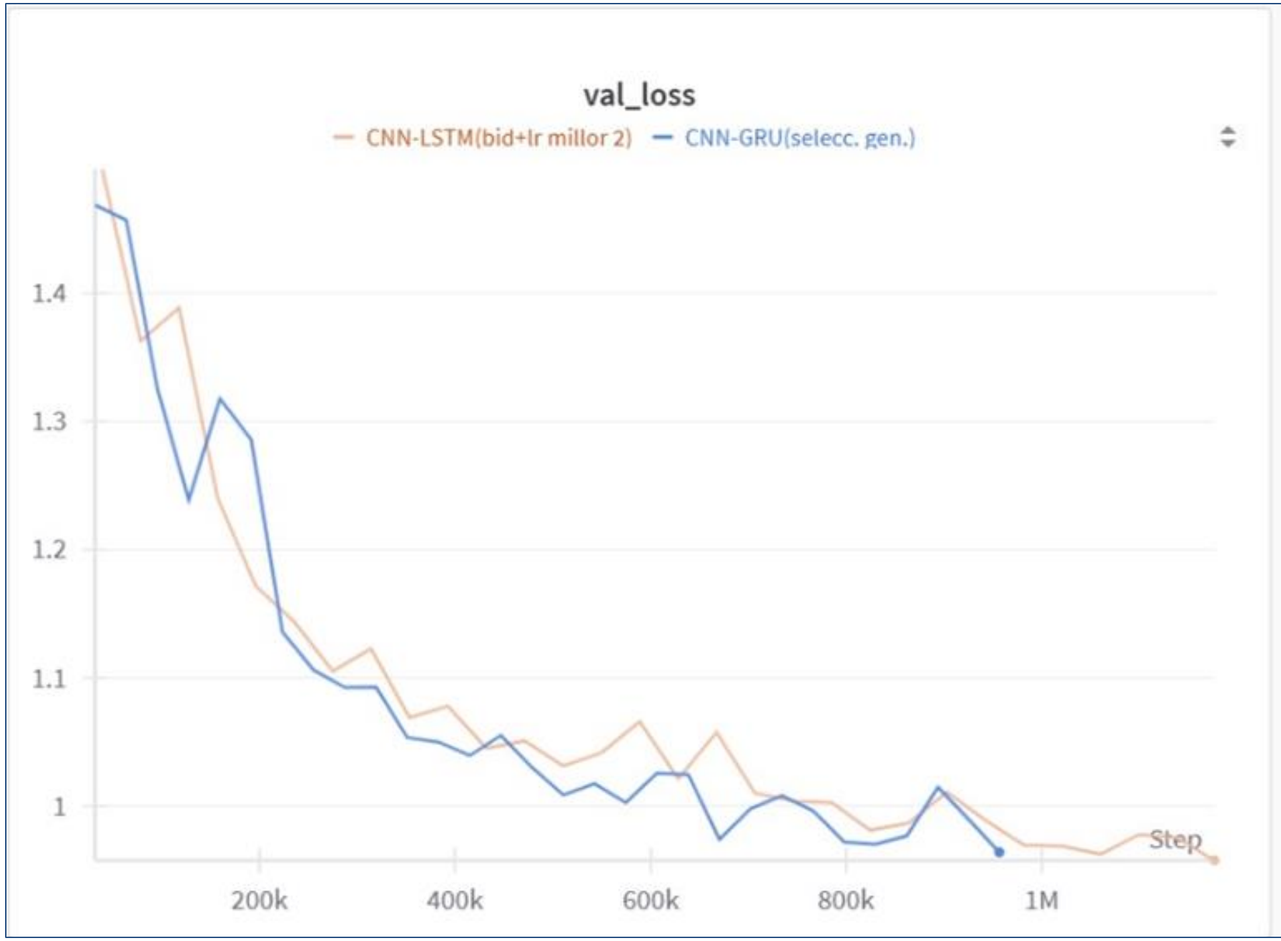
- ReduceLROnPlateau
- Global Pooling
- StepLR





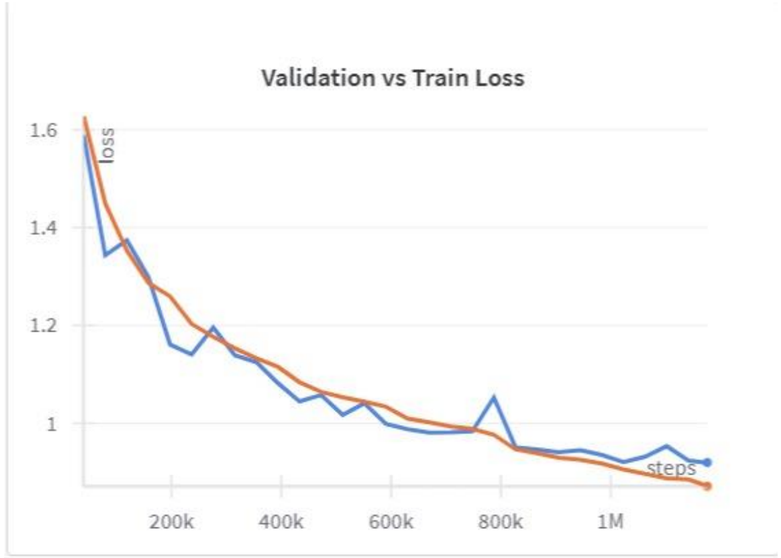

```
reduce_lr_factor: #factor que redueix el LR en el ReduceOnPlateau
    values: [0.5, 0.7]
step_lr_step_size: #número de steps en StepLR
    values: [5, 10]
step_lr_gamma: #factor que redueix el LR en el StepLR
    values: [0.5, 0.7]
```

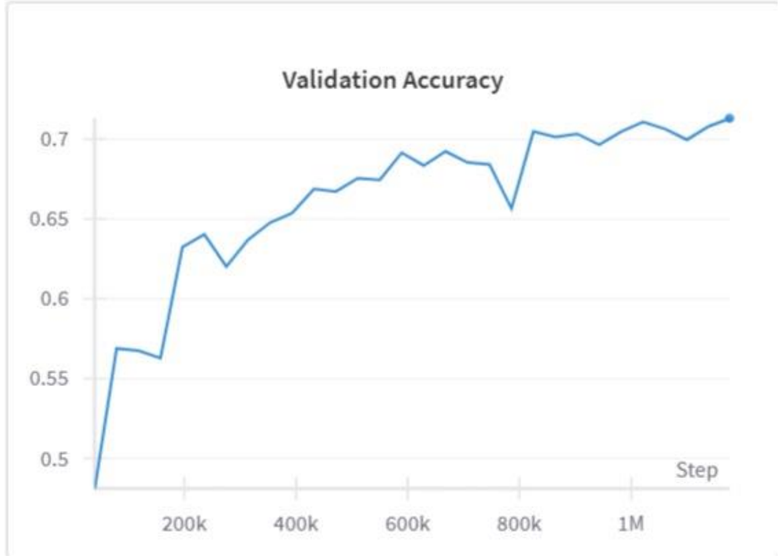
Primers models i millores



CNN(1D) GRU vs LSTM

Primers models i millores

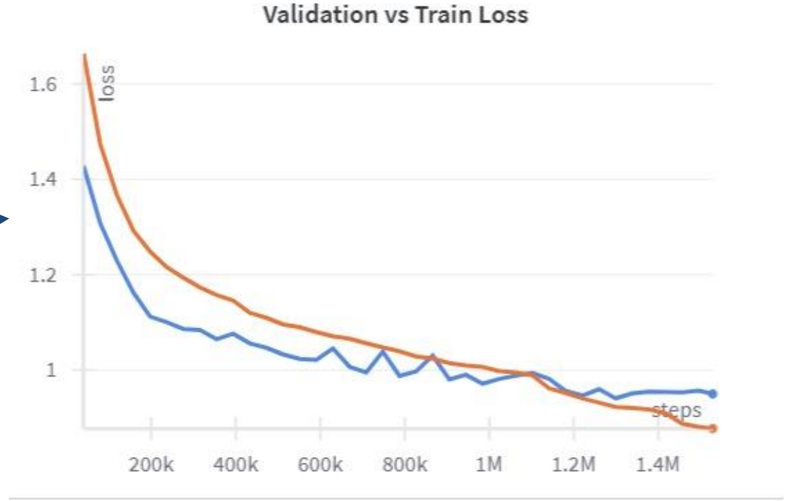


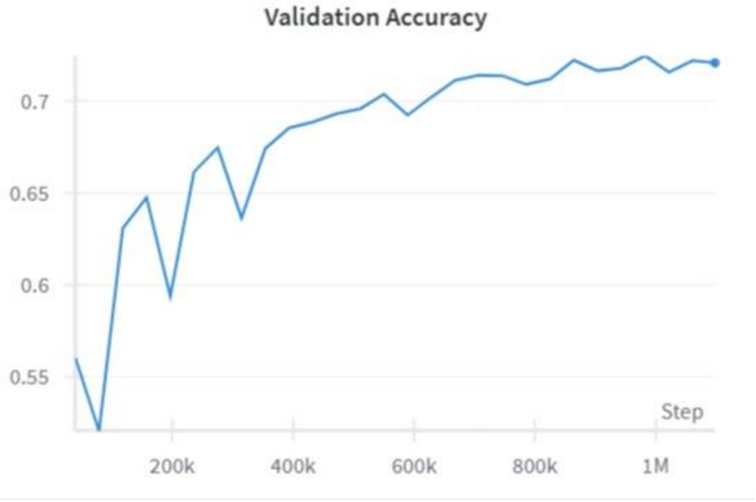


CNN(2D)-LSTM

Primers models i millores

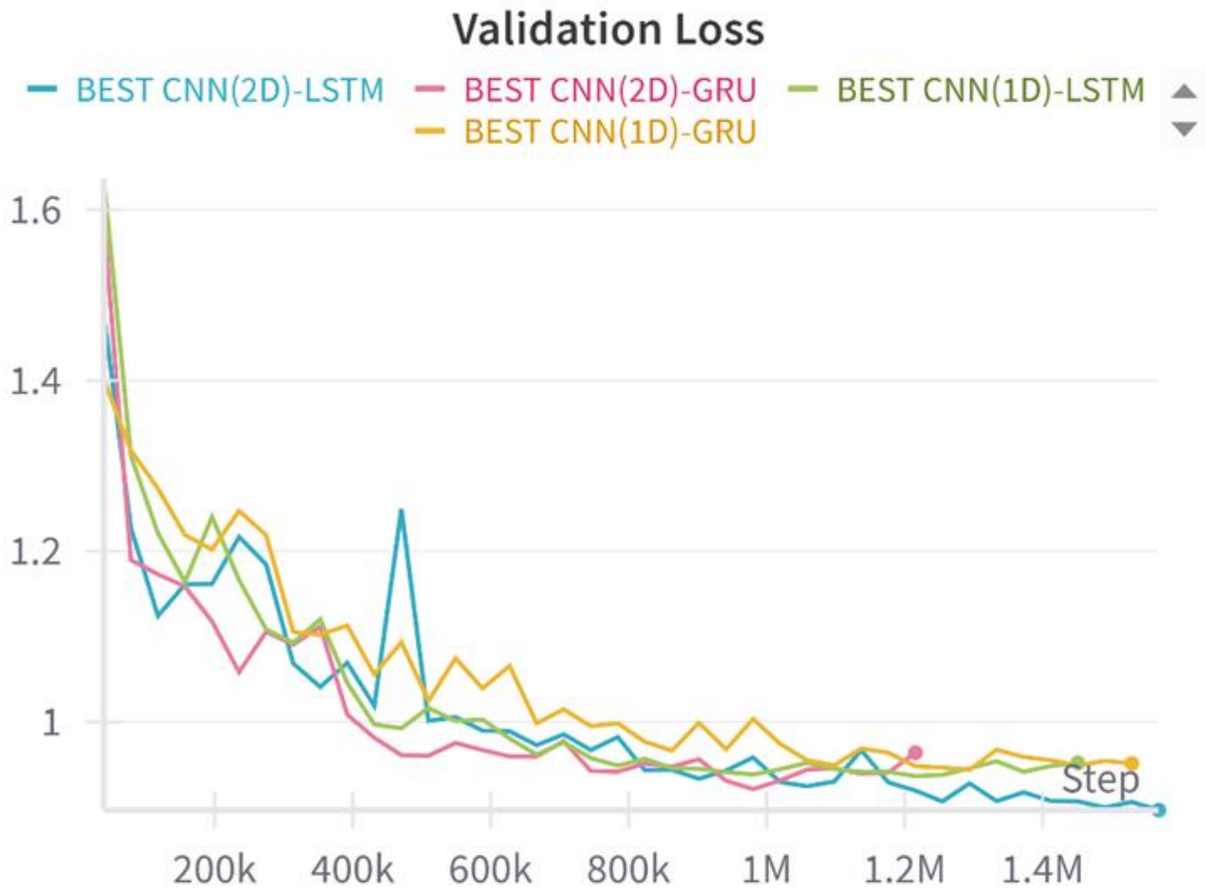
CNN(2D)-GRU

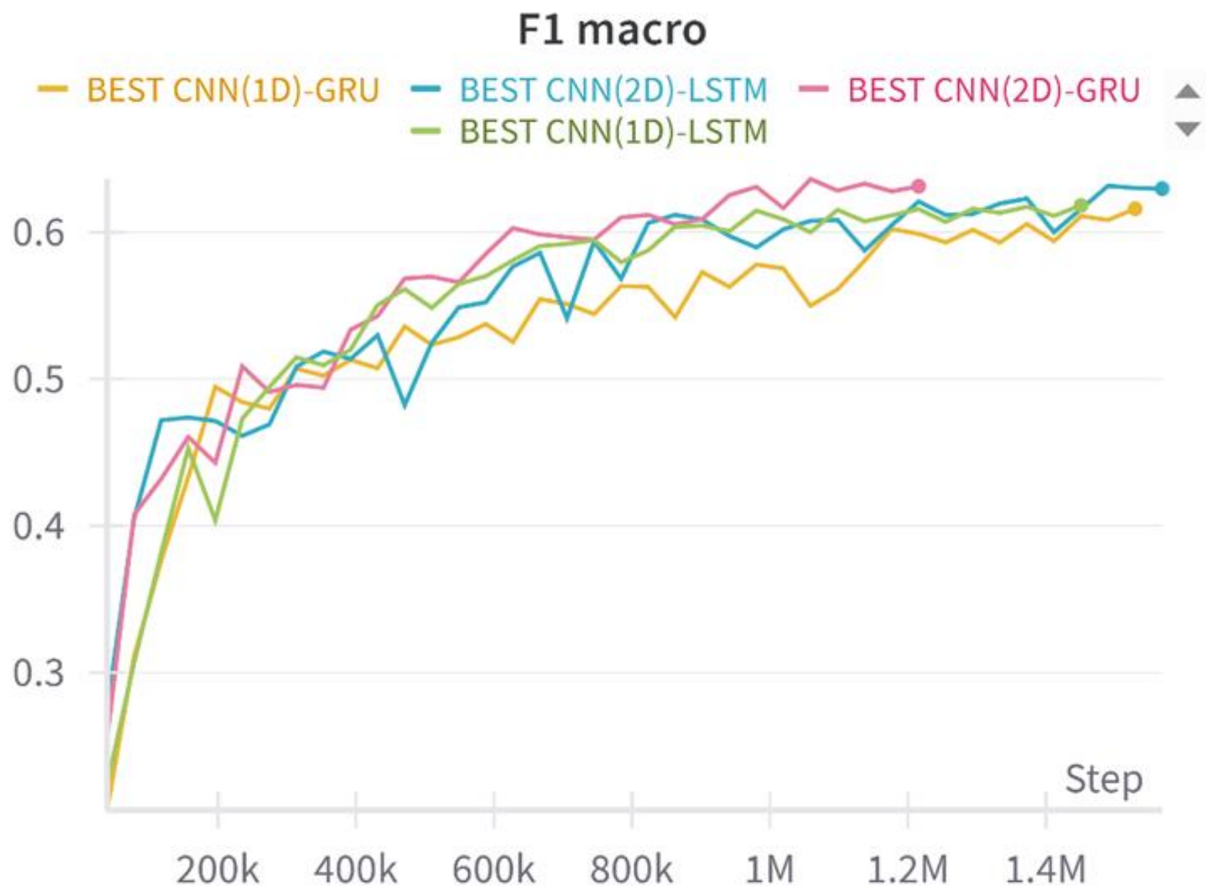




- MaxPool després de les ReLu
- Dropout a la GRU.

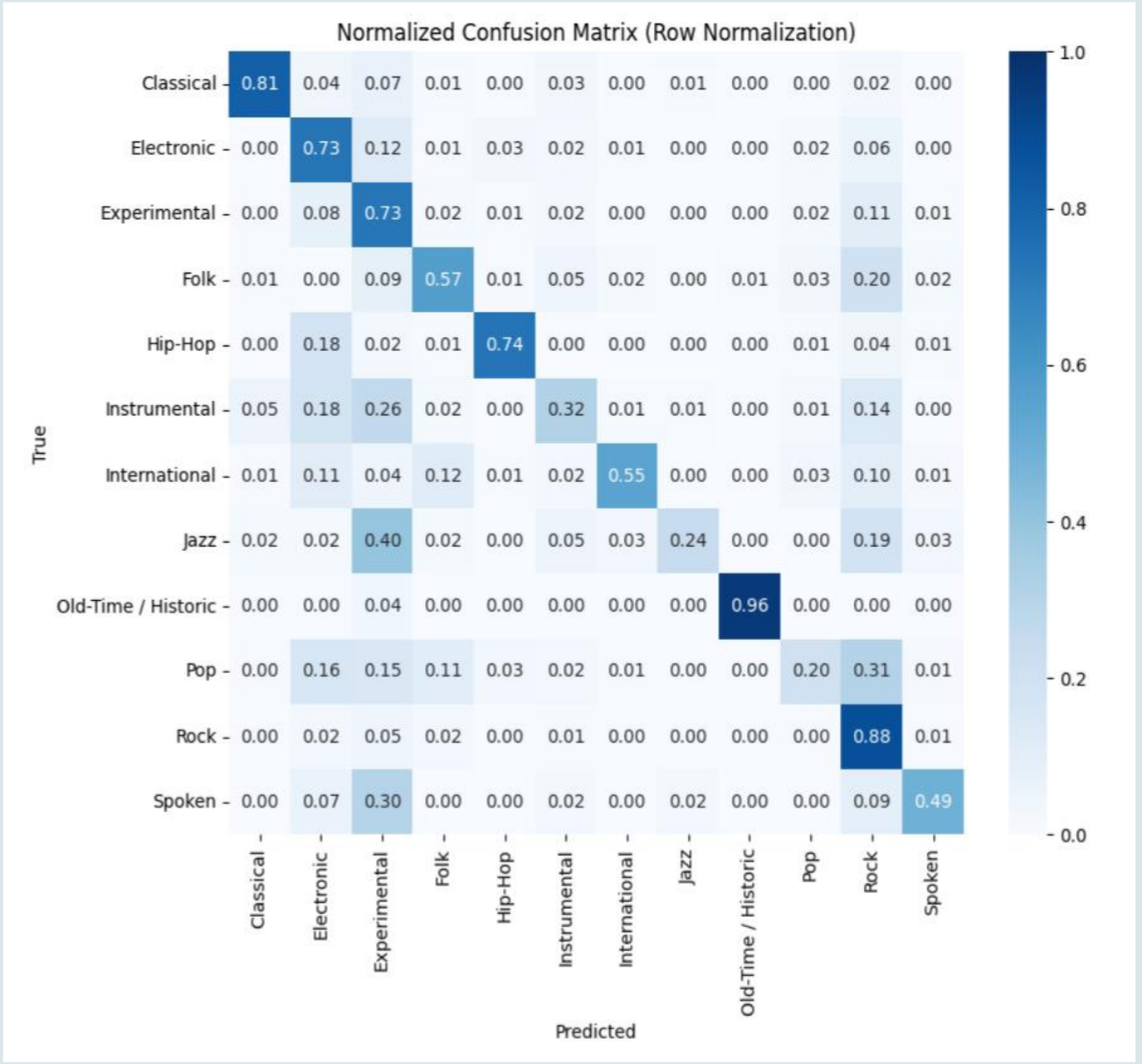
Anàlisi final dels resultats



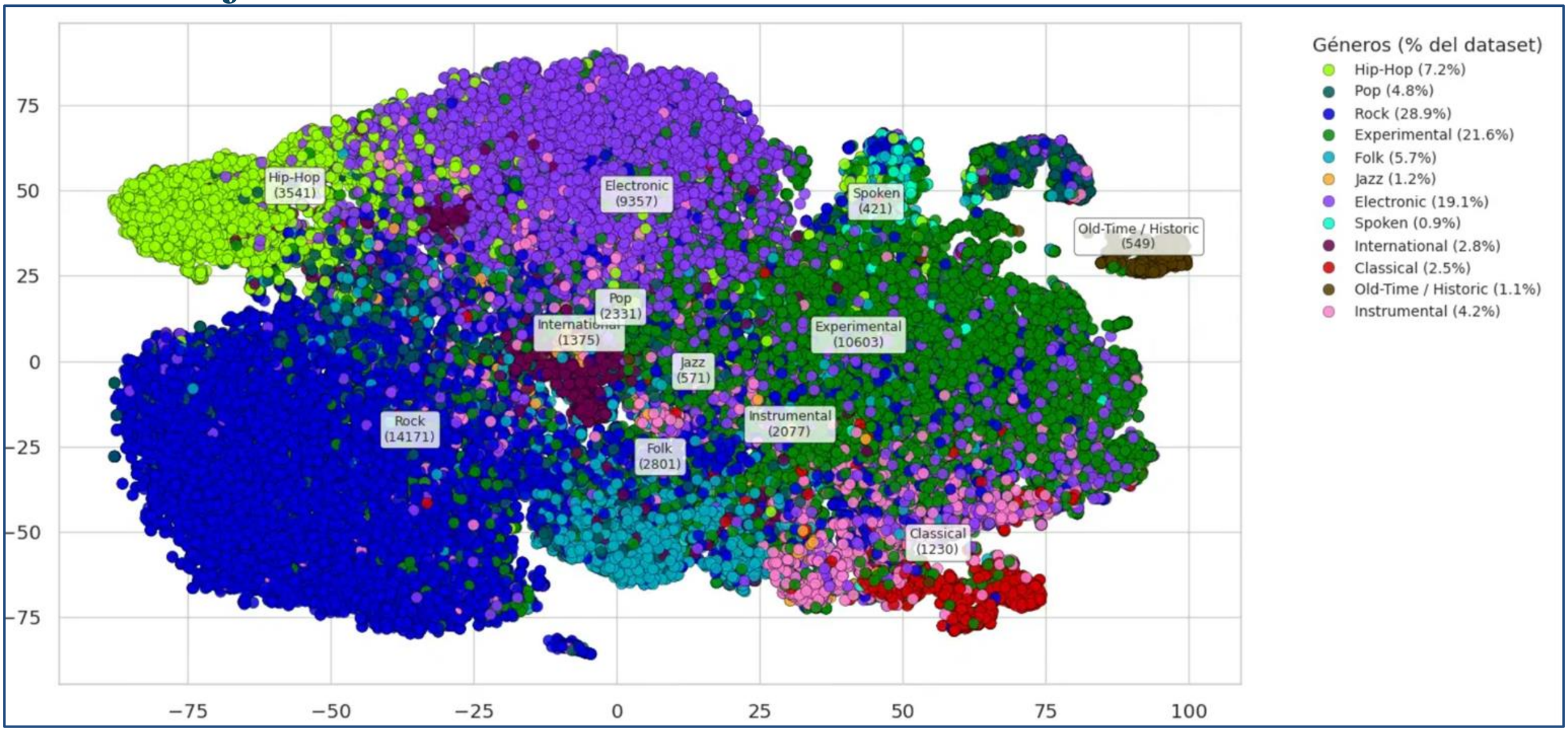


Anàlisi final dels resultats

Arquitectura	Accuracy (%)	F1-weighted (%)	F1-macro (%)
CNN 1D + GRU	69.10	67.40	55.60
CNN 1D + LSTM	69.69	68.64	59.80
CNN 2D + LSTM	70.83	69.70	60.77
CNN 2D + GRU	71.40	70.28	61.90

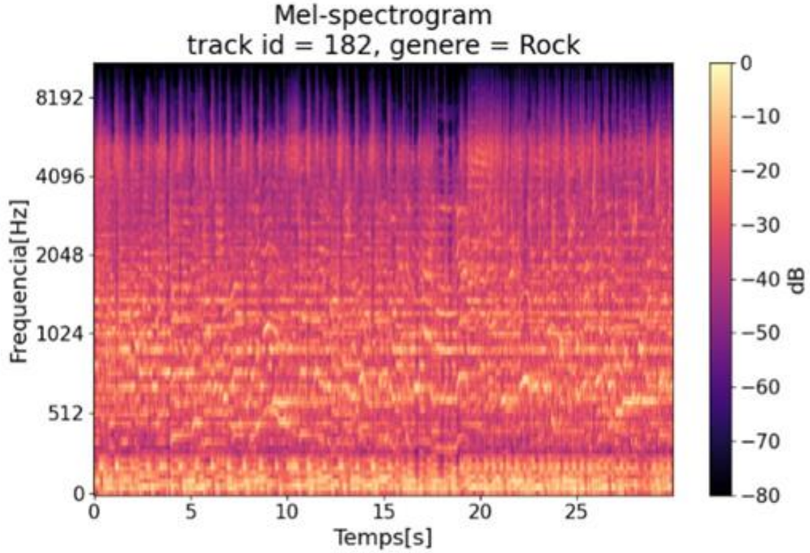

Normalized Confusion Matrix (Row Normalization)
True
Predicted
Classical 0.81 0.04 0.07 0.01 0.00 0.03 0.00 0.01 0.00 0.00 0.02 0.00
Electronic 0.00 0.73 0.12 0.01 0.03 0.02 0.01 0.00 0.00 0.02 0.06 0.00
Experimental 0.00 0.08 0.73 0.02 0.01 0.02 0.00 0.00 0.00 0.02 0.11 0.01
Folk 0.01 0.00 0.09 0.57 0.01 0.05 0.02 0.00 0.01 0.03 0.20 0.02
Hip-Hop 0.00 0.18 0.02 0.01 0.74 0.00 0.00 0.00 0.00 0.01 0.04 0.01
Instrumental 0.05 0.18 0.26 0.02 0.00 0.32 0.01 0.01 0.00 0.01 0.14 0.00
International 0.01 0.11 0.04 0.12 0.01 0.02 0.55 0.00 0.00 0.03 0.10 0.01
Jazz 0.02 0.02 0.40 0.02 0.00 0.05 0.03 0.24 0.00 0.00 0.19 0.03
Old-Time / Historic 0.00 0.00 0.04 0.00 0.00 0.00 0.00 0.00 0.96 0.00 0.00 0.00
Pop 0.00 0.16 0.15 0.11 0.03 0.02 0.01 0.00 0.00 0.20 0.31 0.01
Rock 0.00 0.02 0.05 0.02 0.00 0.01 0.00 0.00 0.00 0.00 0.88 0.01
Spoken 0.00 0.07 0.30 0.00 0.00 0.02 0.00 0.02 0.00 0.00 0.09 0.49
Classical
Electronic
Experimental
Folk
Hip-Hop
Instrumental
International
Jazz
Old-Time / Historic
Pop
Rock
Spoken
1.0
0.8
0.6
0.4
0.2
0.0

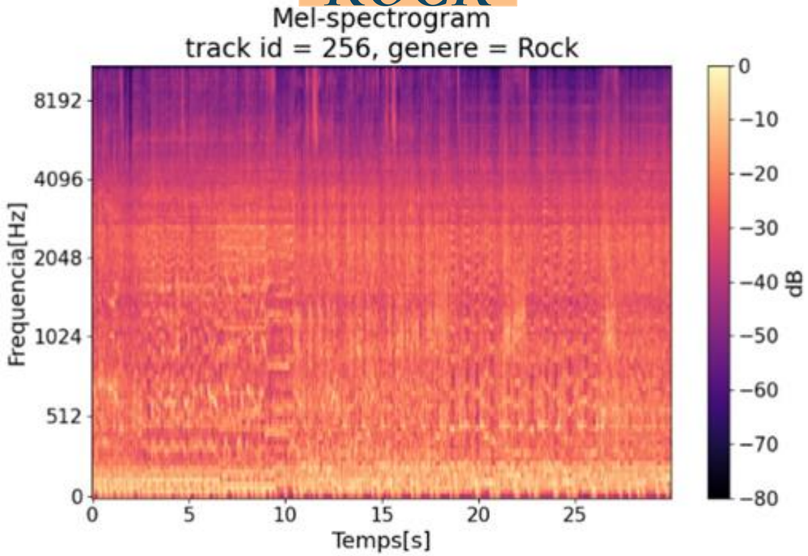
Anàlisi final dels resultats: T-SNE

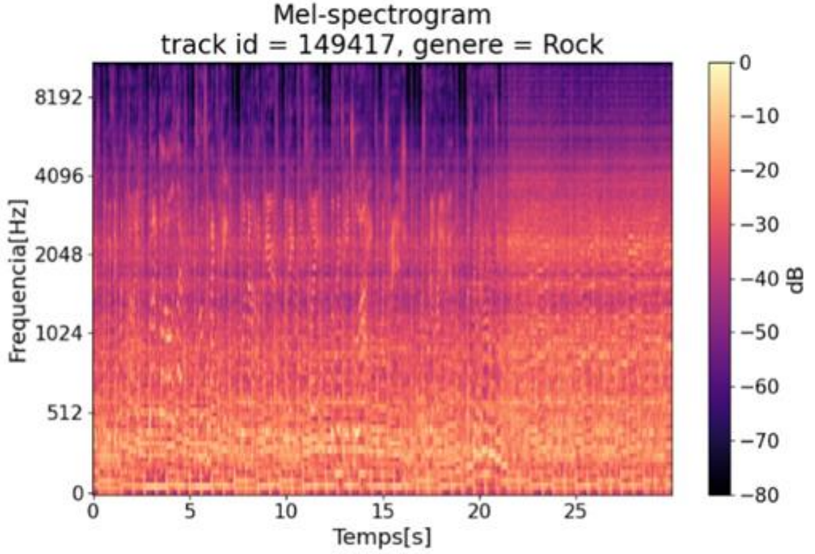


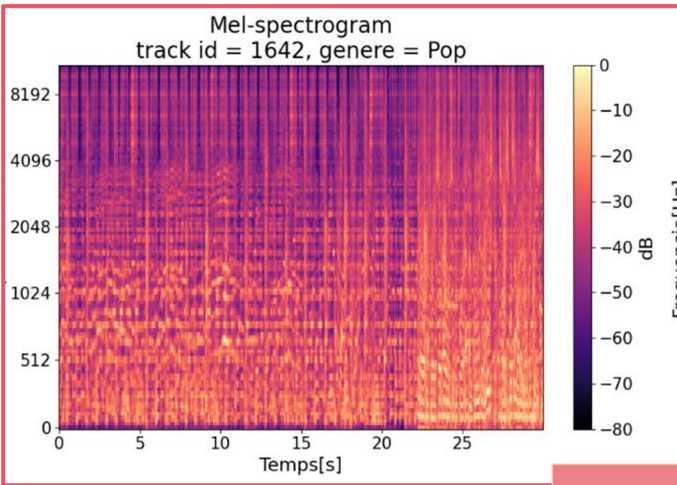
Anàlisi final dels resultats: Visualització mel-spectrogramas

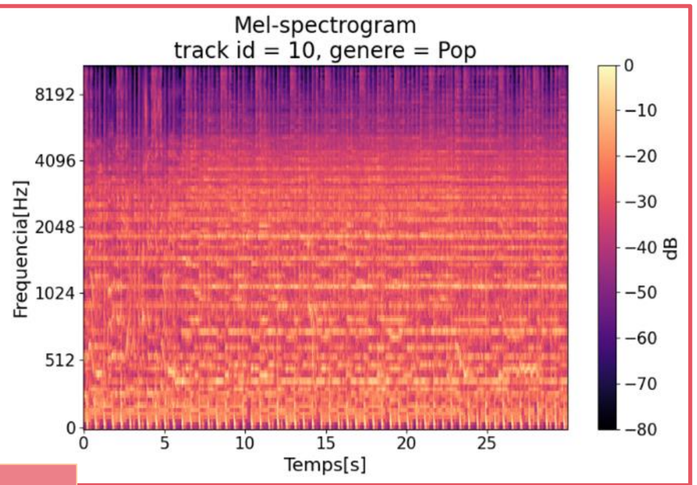
ROCK



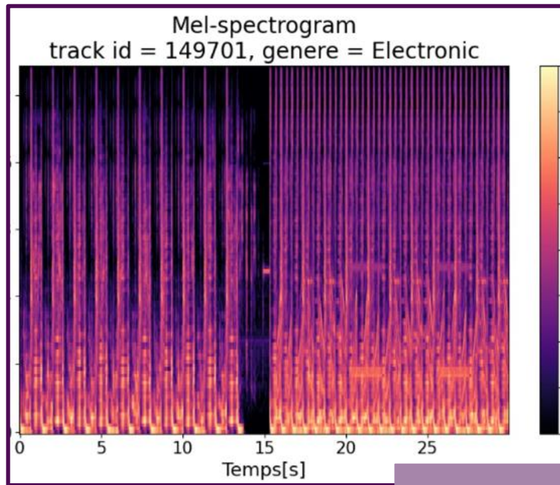


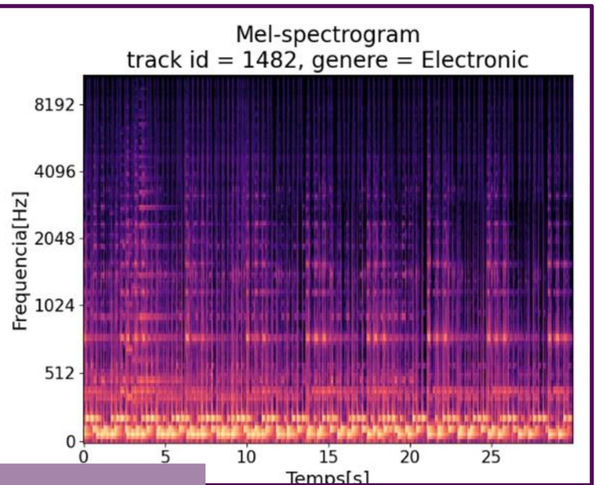






POP





ELECTRONIC

Conclusions