Instrument Activation Detection with CRNN

Ioannis Koutsoukis
MSc in Artificial Intelligence

DEMOKRITOS

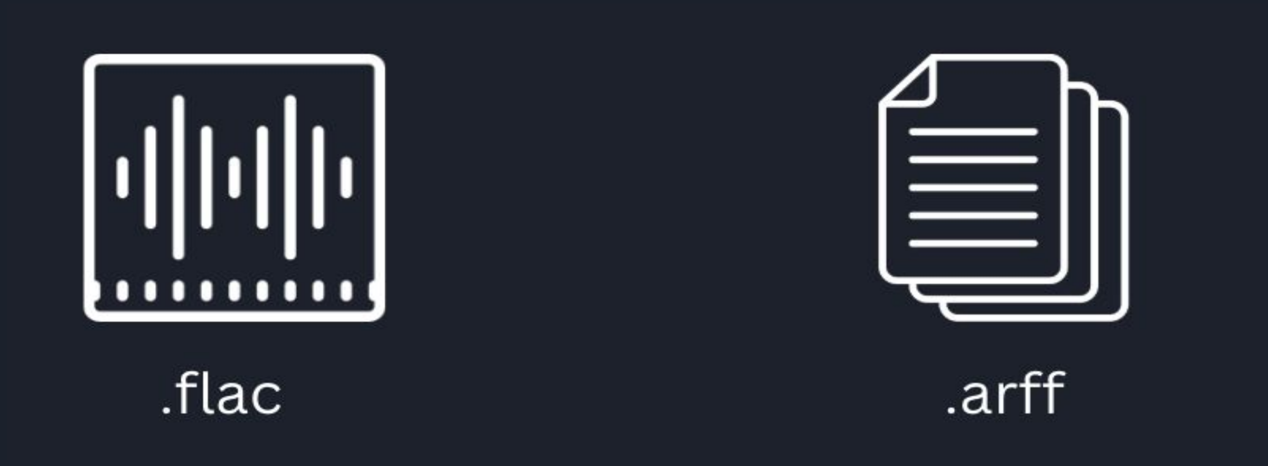
Contents

Problem & Motivation

- **The Objective:** Frame-level multi-label instrument activation detection
- **The Challenge:** Overlapping timbres, limited data, and class imbalance
- **The Value:** Enabling Music Information Retrieval (MIR), source separation, and deep musical content understanding

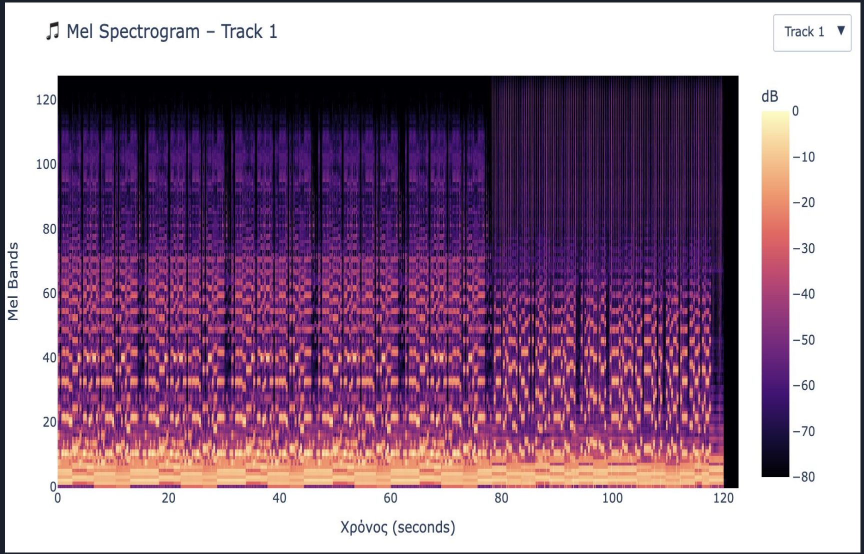
Dataset Overview

- AAM: Artificial Audio Multitracks dataset
    - 3,000 synthesized tracks with isolated stems
    - Each track is accompanied by

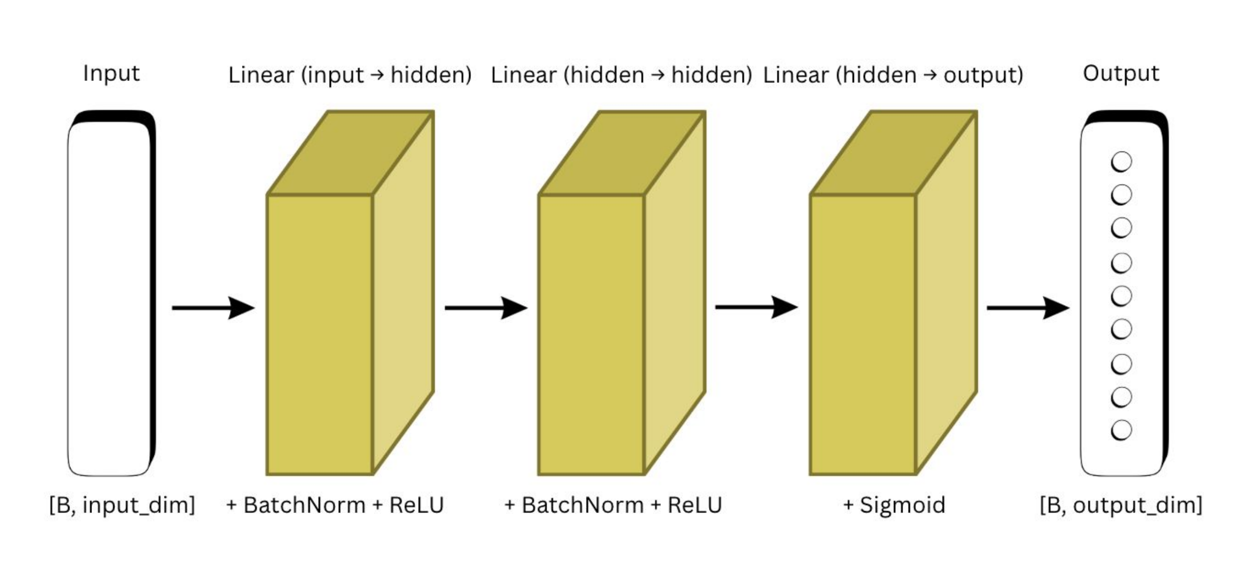


Feature Extraction – Mel Spectrogram (Librosa)

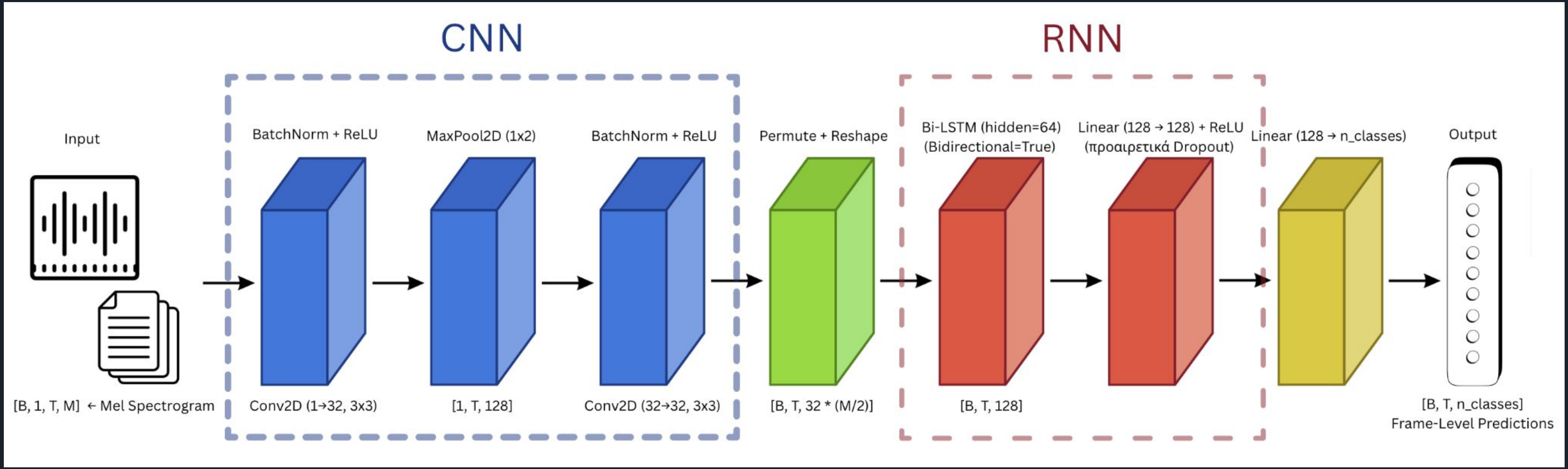
- FFT Window Size (n_fft) = 2048
- Hop Length = 512
- Mel Filters = 128
- Converted to log-scale (dB)



Baseline Model – MLP (Multi-Layer Perceptron)



Proposed Model – CRNN



Training Strategy & Class Imbalance Handling

- Batch Size: 2
- Epochs: 50
- Loss Function: BCEWithLogitsLoss
- Optimizer: Adam
- Frame-level Masking & Padding

- Computed class frequency over training set
  - Rare classes (<5 tracks) were oversampled ×4
  - experimented with class weighting in loss

Evaluation Metrics

- F1 Score (macro / micro / weighted)
  - Macro: treats all classes equally
  - Micro: focuses on global accuracy
  - Weighted: adjusts for class imbalance
- Hamming Loss: average number of wrong labels per sample
- Jaccard Score (sample-based): measures label overlap
- Classification Report per Instrument: per-instrument precision, recall, F1

Evaluation Metrics (Aggr)

```
🎯 Baseline MLP Performance:
F1 Macro:      0.50
F1 Micro:      0.50
F1 Weighted:   0.50
Hamming Loss:  0.51
```

MLP

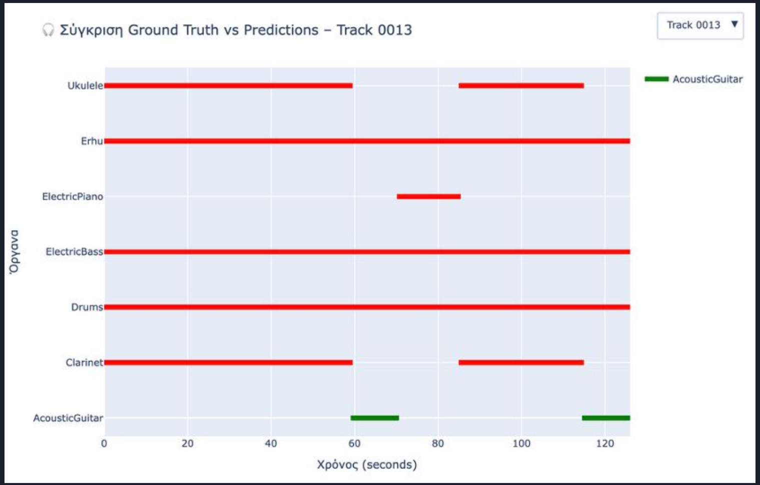
```
🎯 Evaluation Results:
F1 Macro:      0.253
F1 Micro:      0.535
F1 Weighted:   0.537
Hamming Loss:  0.122
Jaccard Avg:   0.387
```

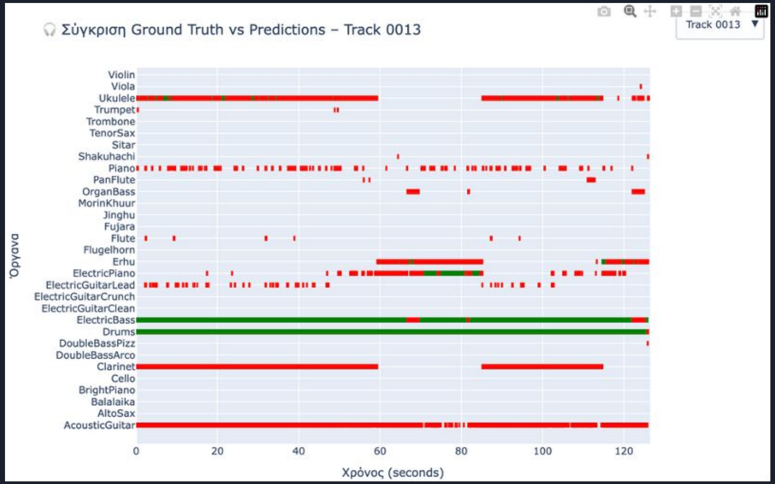
CRNN

Visualizations & Confusion Analysis

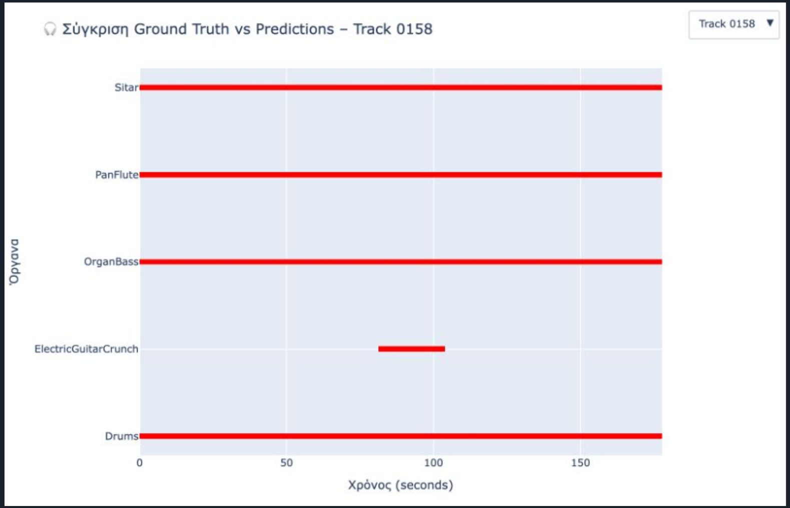
Time to look under the hood...

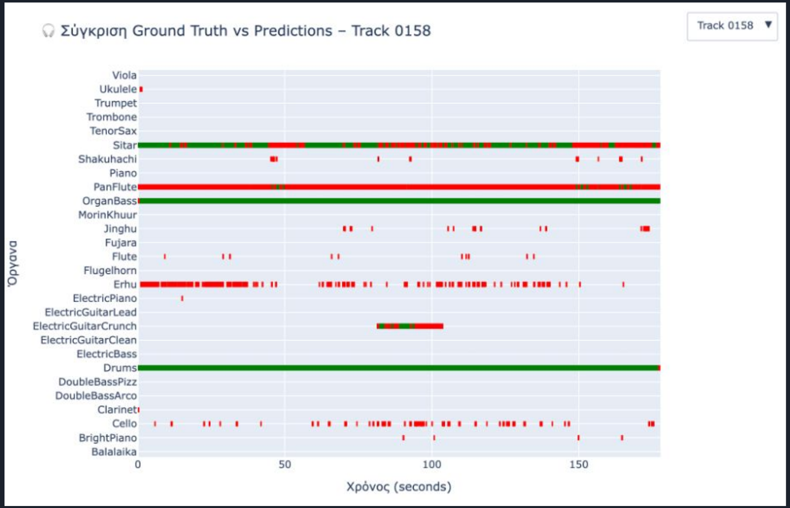
MLP vs CRNN - 150 / 50





MLP vs CRNN - 300 / 50





More Data → Less Confusion

True	Predicted	Count
ElectricGuitarClean	ElectricPiano	11791
Shakuhachi	ElectricPiano	10314
Piano	ElectricPiano	9348
ElectricGuitarClean	Ukulele	7641
Erhu	Fujara	7050
Erhu	Sitar	6974
TenorSax	ElectricPiano	6645
Shakuhachi	Ukulele	6614
TenorSax	Piano	6289
Piano	Ukulele	5899
Flugelhorn	ElectricBass	5445
Jinghu	Sitar	5115
AcousticGuitar	ElectricPiano	4975
BrightPiano	ElectricPiano	4930
ElectricGuitarClean	Piano	4896

CRNN - 150 Songs

	True	Predicted	Count
3	Erhu	Cello	1650
10	Erhu	Flute	1133
8	Erhu	Jinghu	546
15	Erhu	Drums	323
2	DoubleBassPizz	Cello	218
22	Sitar	Cello	136
16	Erhu	Trumpet	127
19	Erhu	Balalaika	86
7	Drums	Jinghu	82
6	DoubleBassPizz	Jinghu	68
20	Sitar	Balalaika	59
11	Drums	Flute	57
1	Erhu	Clarinet	45
9	Drums	Cello	41
12	Erhu	BrightPiano	37

CRNN - 300 Songs

Conclusions

- MLP performs reasonably well in low-data scenarios (based on metrics) or when frame-level is not required
- CRNN performs worse at low data but scales better with more data (based on metrics)
- Frequent confusions are observed in overlapping spectral instruments (e.g., flute vs clarinet)
- Weighted losses + oversampling help address class imbalance

Future Work

- Add more training data from the full AAM set
- Experiment with pre-trained CNNs for better audio embeddings
- Evaluate attention-based and Transformer-based models
- Study temporal consistency with post-processing smoothing

Citations

- Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP.2017.7952585
- Huang, Y., Wang, J., & Liu, Y. (2022). Audio Spectrogram Transformer: Learning on the Time-Frequency Representation of Audio. arXiv preprint arXiv:2104.01778.
- Won, M., Ferraro, D., Han, Y., & Nam, J. (2020). Evaluation of CNN-based automatic music tagging models. arXiv preprint arXiv:2006.00751.

Any Questions?

Thank you for your Time