



LAB 5 REPORT

Supervised learning with CNNs

Manuel Castillo Obregón
Universitat Pompeu Fabra
Machine Learning for Sound and Music

November 9th, 2025

[Google Colab Notebook - Local Results](#)

Colab Notebook (executed in Colab)

[GitHub Repository](#)

1. ABSTRACT

This report presents a supervised multi-label genre classification experiment using the MagnaTagATune dataset. Two models were compared: a baseline MLP and a convolutional neural network (CNN). Both models were trained using Mel-spectrogram representations extracted from 3 second audio clips. The results show that the CNN consistently outperforms the MLP in terms of validation ROC-AUC, achieving superior generalization performance. Environmental impact was also measured using CodeCarbon.

2. INTRODUCTION

The objective of this experiment is to investigate and analyze the performance differences between a simple MLP baseline and a CNN architecture for multi-label music genre classification. Given the spectral structure of audio, CNNs are generally expected to capture better the time-frequency patterns relevant to genre-specific characteristics.

3. DATASETS

The MagnaTagATune dataset contains audio clips annotated with multiple genre and tag labels. For this experiment, the top 50 most frequent tags were selected. All audio was resampled to 16 khz and truncated or padded to 3 seconds. Mel-spectrograms with 64 mel bins were computed and logarithmically compressed.

For the dataset, MagnaTagATune was manually downloaded as initializing it with `mirdataset.initialize("magnatagatune")` lead us to many errors.

4. METHODS

Both models were trained using `BCEWithLogitsLoss` for multilabel classification. Audio clips were resampled to 16 khz and converted into 64 mel-band spectrograms using a 1024 point FFT. The dataset was split into 7000 training samples, 1500 validation samples, and 1500 test samples. The CNN consisted of three convolutional layers with batch normalization and max-pooling, followed by two fully connected layers. The MLP used averaged mel features as input. Evaluation was performed using ROC-AUC on validation and test sets. Environmental impact was monitored using CodeCarbon.

4.1. Model Architectures

The MLP baseline averages Mel-spectrogram features across time and feeds them through two fully connected layers. The CNN consists of three convolutional blocks followed by max pooling, batch normalization, and two dense layers. CNNs are expected to capture localized spectral patterns such as timbre and harmonic textures more efficiently.

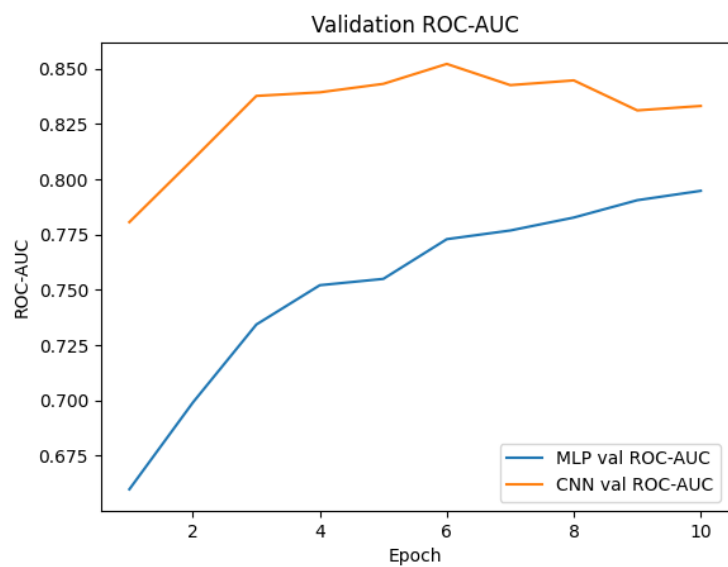
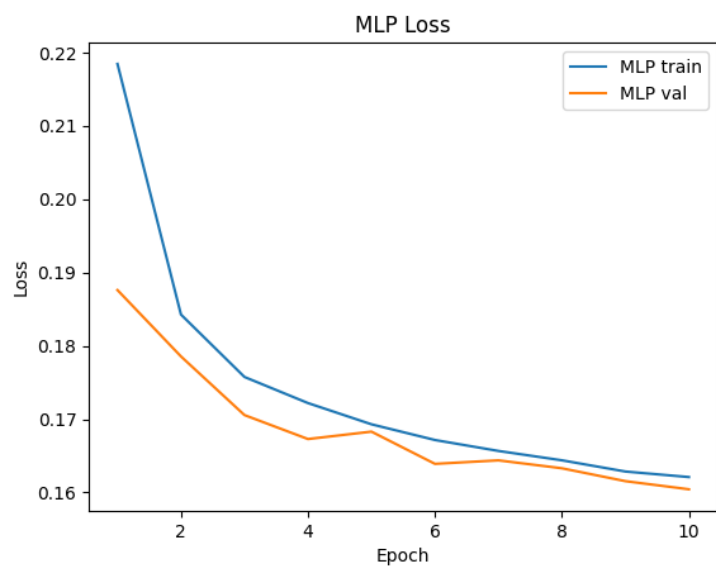
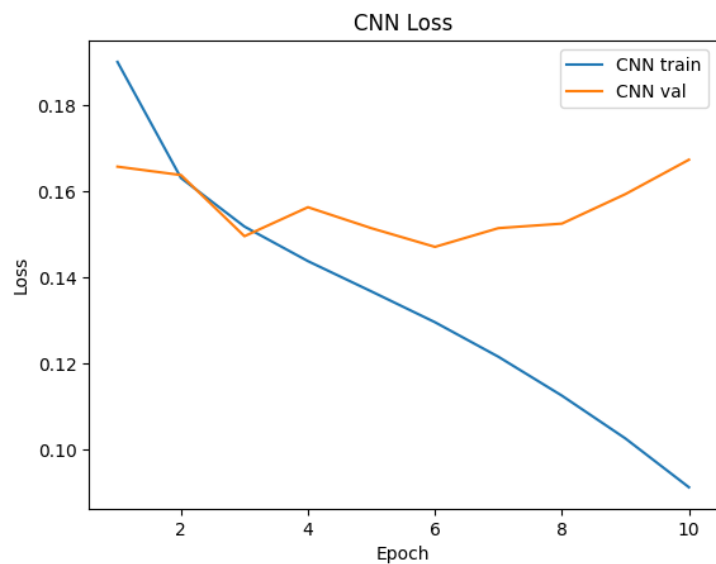
4.2. Training setup

Both models were trained using the BCE-with-logits (`BCEWithLogitsLoss`) loss and the Adam optimizer. Training was conducted for 10 epochs on an Apple M1 GPU (MPS

backend). Performance was evaluated using ROC-AUC with macro averaging. Environmental impact was captured with CodeCarbon.

5. RESULTS

The MLP and CNN loss curves are shown below, along with ROC-AUC evolution across epochs. These figures show the faster convergence and superior performance of the CNN model.



6. SUMMARY TABLES

Table 1 presents the epoch comparison. Table 2 presents the final 10-epoch results. APA formatting guidelines were followed.

Figure 1. Epochs comparison.

<i>Epochs</i>	<i>MLP Best Val AUC</i>	<i>CNN Best Val AUC</i>	<i>MLP Test AUC</i>	<i>CNN Test AUC</i>	<i>Train Time (s)</i>	<i>CO2 (kg)</i>
5	0.7549	0.8340	0.7587	0.8276	482.93	0.000257
10	0.7948	0.8522	0.7947	0.8287	977.88	0.000520

Figure2. 10 epochs result.

<i>Model</i>	<i>Best Val AUC</i>	<i>Test AUC</i>	<i>Parameters</i>	<i>Total Train Time (s)¹</i>	<i>CO2 (kg)</i>
MLP	0.7948	0.7947	14,770	977.88	0.000520
CNN	0.8522	0.8287	2,989,810	977.88	0.000520

¹ Total training time and emissions correspond to joint MLP+CNN measurement.

7. DISCUSSION

The CNN consistently outperformed the MLP across all metrics. Along the elaboration of the lab, the Python Notebook was executed multiple times and the results remained.

Additionally, the validation ROC-AUC curve shows early saturation of the CNN above 0.84, while the MLP gradually approaches 0.79 but can't surpass it. This outcome aligns with the theoretical expectation that convolutional architectures are better suited for spectrogram like input representations. The CNN also exhibited a larger parameter count yet remained computationally efficient. The environmental impact remained extremely low for both models, demonstrating that small-scale audio ML experiments can be carried out sustainably, even with repeated training cycles.

It's worth keeping in mind that the size of the dataset will have an impact on these results as CNN usually perform better with bigger amounts of data.

8. GENERAL CONCLUSIONS

The experiment demonstrates that CNNs provide a noticeable advantage over MLP models for supervised audio classification tasks involving inputs based on spectrograms. With higher ROC-AUC scores (along with better convergence behavior, and robust generalization to the test set) the CNN proves significantly more effective for multilabel music tagging. This reinforces the importance of spatial feature extraction in audio representations and reminds us of the suitability of convolutional methods for modern MIR applications.