# Data Augmentation for Environmental Sound Classification - How to Deal with Small Datasets

MLHD - 2024/25
Federico Chiarello - ID: 2058163

# Project Objectives

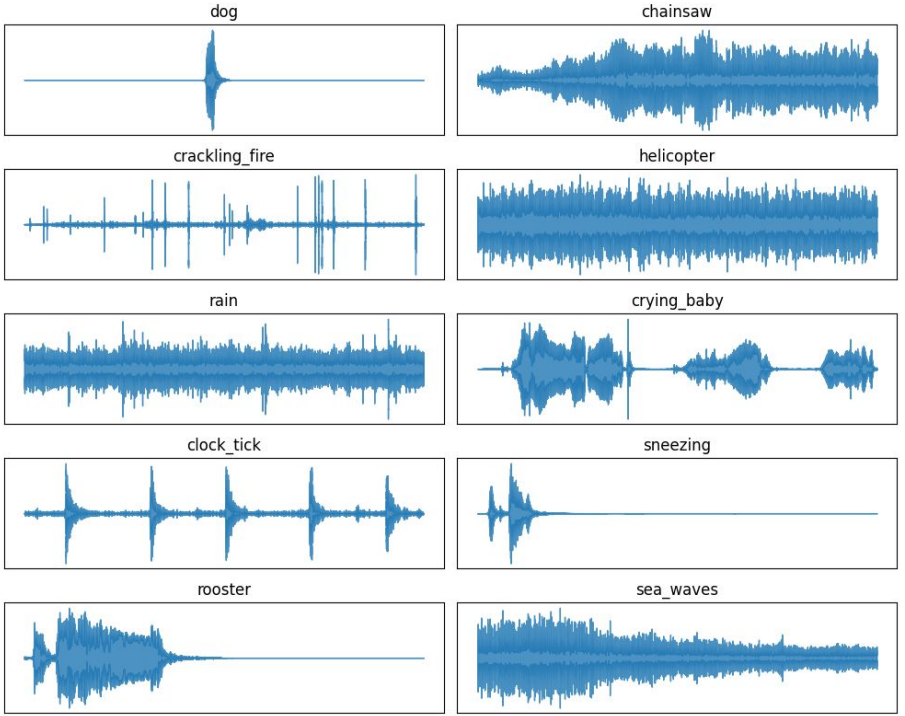
- Perform **Environmental Sound Classification** on ESC-10 dataset
- Test the potential of **Audio-Augmentation** Techniques
- Implementation of **CNN** models to perform classification on **Mel Spectrograms**

# ESC-10 Dataset
## and
# Preprocessing Pipeline

# Dataset

**ESC-10** Dataset:

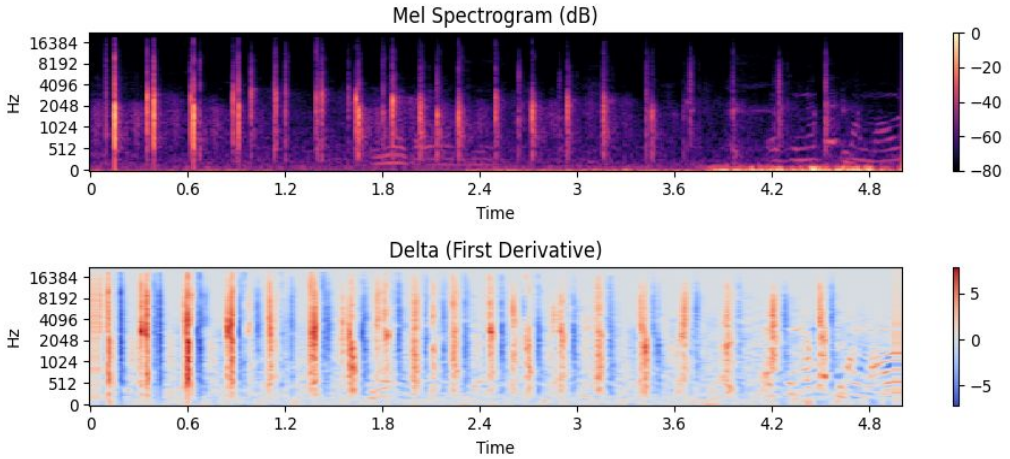
- 5 seconds audio samples
- 10 balanced classes
- 400 samples
- 5 CV-folds



# Log Mel-Spectrograms

**Audio Processing:**

- **Sampling Rate**: 22050 Hz
- 60 mel-bends
- First temporal derivative
- Two-dimensional Feature Map



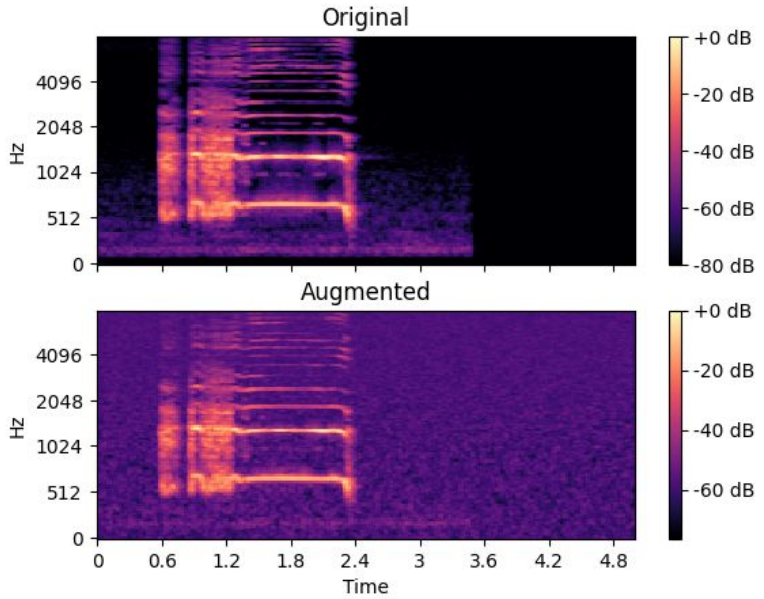
# Data Preprocessing Pipeline

**TensorFlow Dataset Pipeline**:

1. Data Loading
2. **Caching**
3. Shuffling
4. Repeating
5. **Audio Augmentation**
6. Computing Mel-Spectrograms and Deltas
7. Normalizing
8. Batching
9. **Prefetching**

# Data Augmentation

- Time Stretching
- Pitch Shifting
- Gaussian Noise
- Audio Shifting



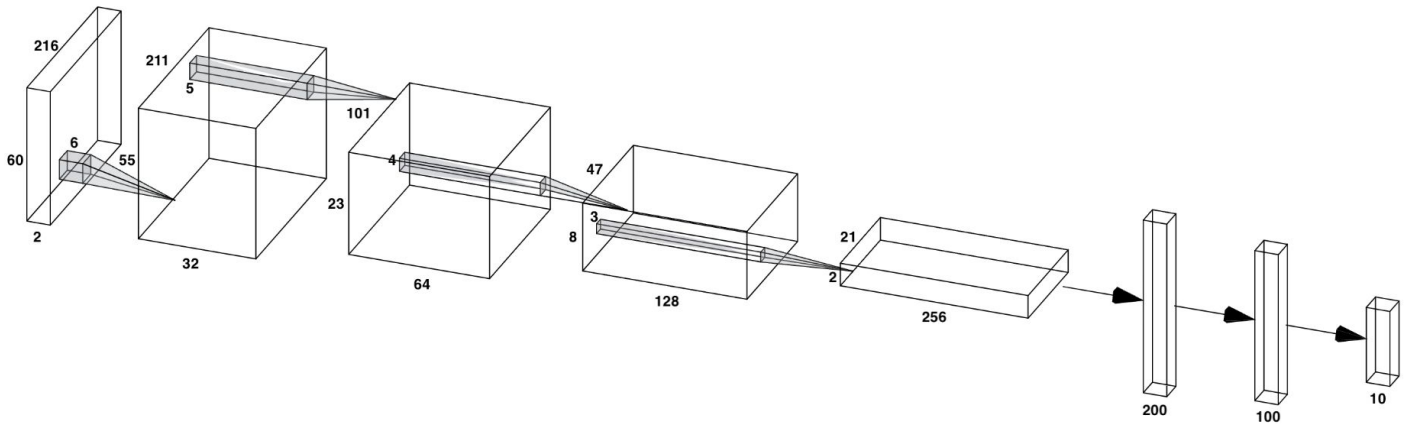
# Audio Augmentation Strategies

| **Audio Augmentation** | Schema **A1** | Schema **A2** |
|---|---|---|
| Time Stretching | (0.75, 1.25) | (0.5, 1.5) |
| Pitch Shifting | 2 | 4 |
| Gaussian Noise | 0.01 | 0.02 |
| Max Shifting | 0.2 | 0.4 |
| Padding | 0.002 | 0.004 |

# CNN Models and Results

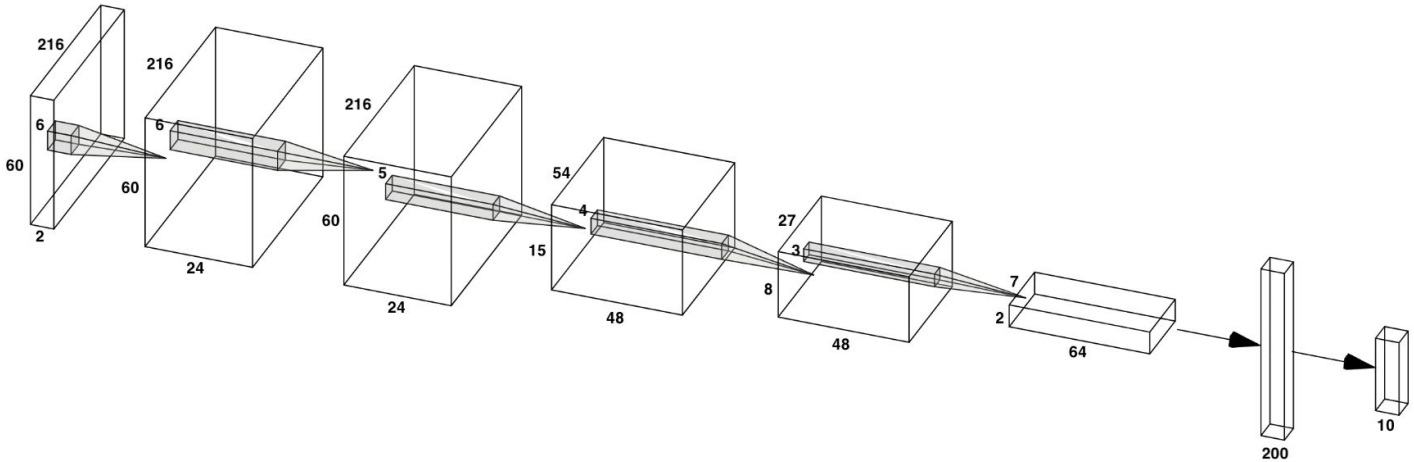
# Baseline CNN

- Increasing **filters** count: (32, 64, 128, 256)
- Decreasing **kernel** size: (6, 5, 4, 3)
- **Max-Pooling** layers
- **Dropout** before fully-connected layers
- **1M** Parameters



# MelNet CNN

- **Batch Normalization** layer before fully-connected
- **Dropout** layers to mitigate **overfitting**
- **Max-Pooling** layers to reduce spatial dimensionality
- **340k** Parameters



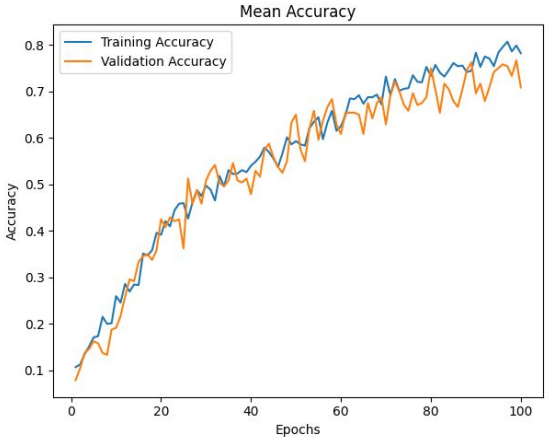
# Learning Framework

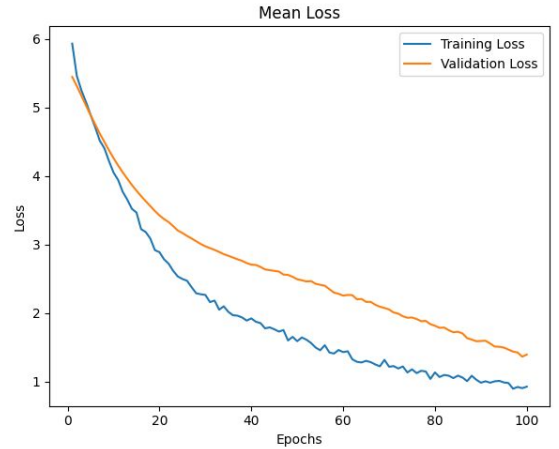
- 100/150 epochs with **Early Stopping**
- **Adam** Optimizer
- Sparse Categorical Crossentropy Loss

- 3 folds for **Training**, 1 for **Validation**, 1 for **Testing**
- **Batch Size** of 80

# Results

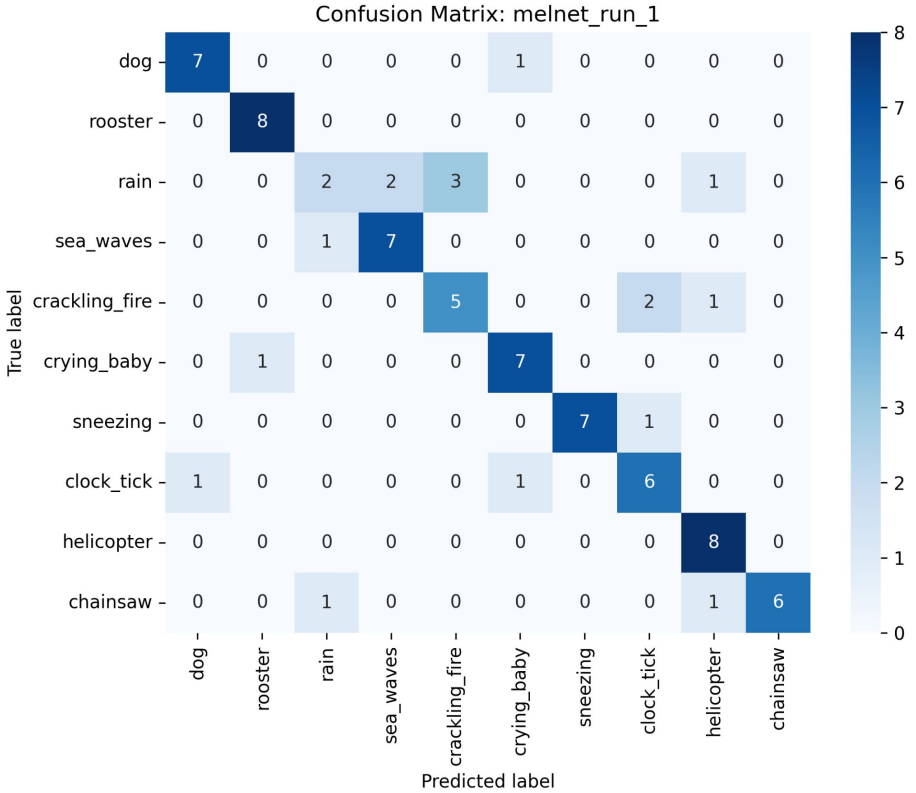
| Model | Data Augmentation | Test Accuracy |
|---|---|---|
| Baseline | None | 61.2 |
| Baseline | A1 | **66.3** |
| Baseline | A2 | 61.2 |
| MelNet | None | 71.2 |
| MelNet | A1 | 68.8 |
| MelNet | A2 | **75.0** |



# Conclusions

- **Audio Augmentation** could significantly improve generalization
- Some **classes** proved to be particularly challenging for the model



# Demo