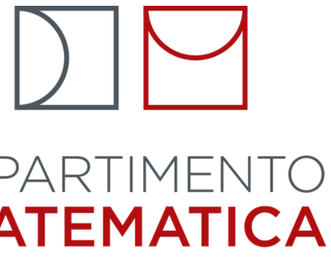
A comparative study of input features and augmentation techniques in CNN-CRNN models for Enviromental Sound Classification

Human Data Analytics, A.Y 2023/2024

Laura Legrottaglie
ID: 2073222



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Introduction

Environmental sounds encompass a wide range of everyday audio events that cannot be classified as either speech or music

Dog barking

Glass breaking

Clapping hands

Introduction

Environmental sound classification (ESC) has received increasing research attention due to its **applications**:

Introduction

Challenges: ?

- Limited data
- Variability across different sounds
- Lack of standard methodology and evaluation

Machine and Deep Learning
models offer **potential solutions**

A general overview

We aim to answer to the following questions:

- How **different model architectures** can solve the ESC task?
- How **different types of feature representations** affect the predictions?
- What are **the most effective augmentation techniques** to use?

The dataset

Università degli Studi di Padova

The performance of the proposed models is evaluated using the **ESC-50 dataset,** developed by **Piczak**, the first one to use CNN models to solve the ESC task.

Key characteristics: 2000 audio samples, 50 classes

Animal sounds 🐾

Dog
Cat
Pig
Cow
Insects
Sheep
Rooster ...

Nature sounds 🌳

Rain
Wind
Water drop
Sea waves
Thunderstorm
Crackling fire
Chiping birds...

Human sounds 👨🏽

Crying baby
Sneezing
Clapping
Footsteps
Laughing
Snoring
Breathing...

Domestic sounds 🏠

Clock alarm
Can opening
Glass breaking
Vacuum cleaner
Washing machine
Clock tick...

Exterior/urban sounds ⚙️

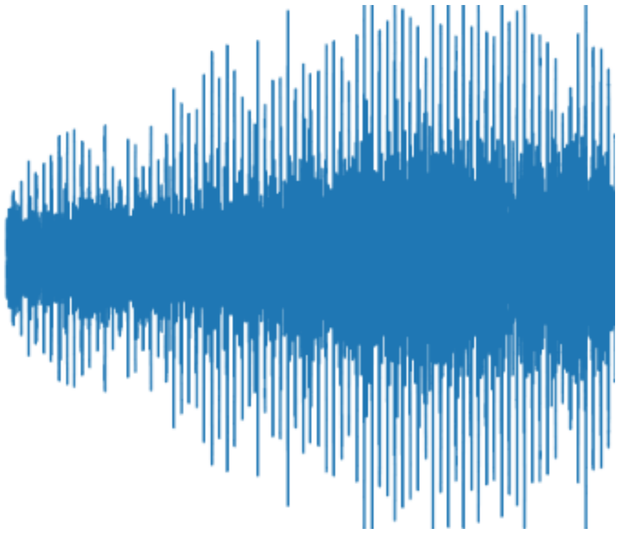
Engine
Chainsaw
Helicopter
Train
Airplane
Siren
Hand saw...

The dataset

Università degli Studi di Padova

Other important aspects:

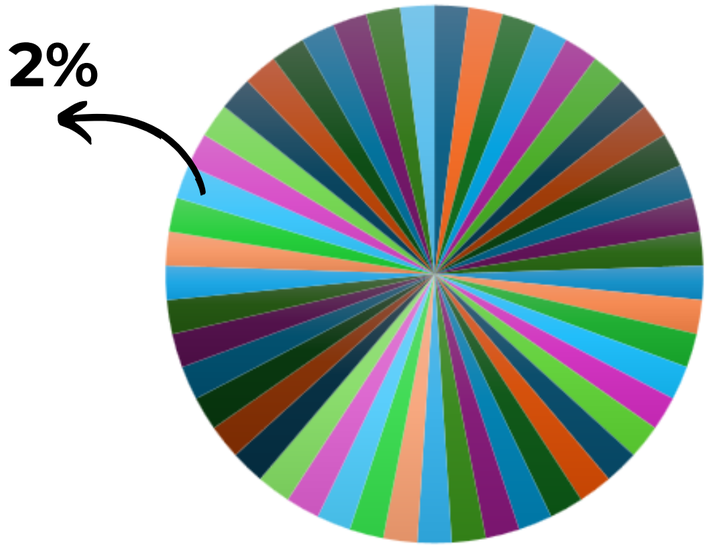
Audio characteristics



5 seconds

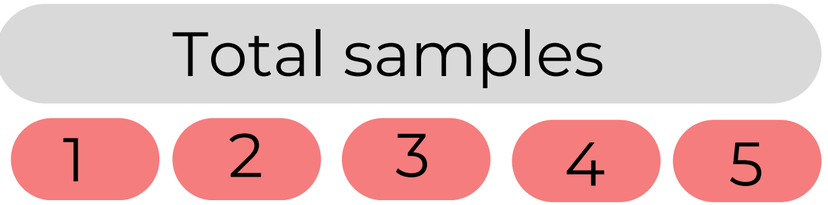
Single channel 44,1 KHz

Class distribution



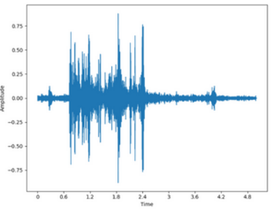
40 clips for every class

5 equal cross-validation folds



Same original audio = same fold

Processing pipeline



White noise injection

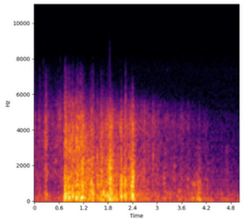
Pitch shifting

Data augmentation

RawCNN

Frog
Siren
Engine
Wind

Feature extraction



Log-mel Spectrogram

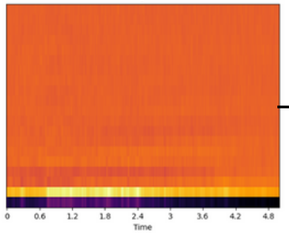
Mixup

Logmel-CNN

Frog
Siren
Engine
Wind

Logmel-CRNN

Frog
Siren
Engine
Wind



MFCC

Mixup

MFCC-CNN

Frog
Siren
Engine
Wind

Audio preprocessing

UNIVERSITÀ DEGLI STUDI DI PADOVA

Key Steps

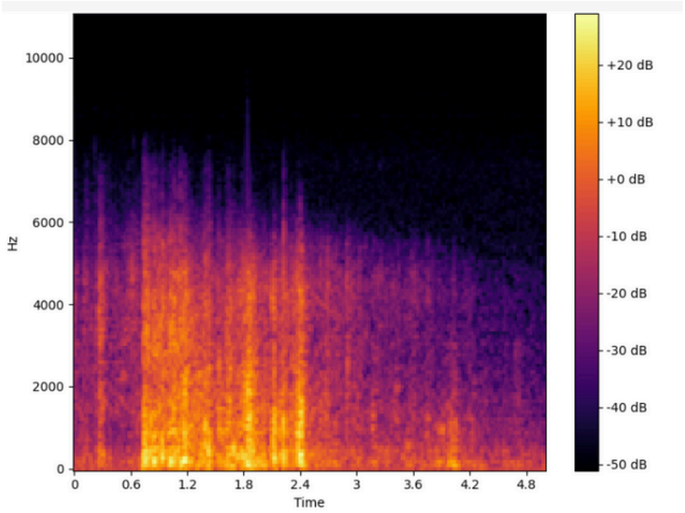
- File Format: .wav format.
- Resampling: 22,050 Hz
- Normalization: Ensured all sample values are scaled to:
  - [-1, 1] for spectrogram and MFCC extraction.
  - [0, 1] for raw waveform model.

Features extraction

Features extraction

Log-Mel Spectrograms

- Window size: 1024, Hop length: 512, Mel bands: 60
- Delta features (rate of change over time) as second channel
- Input representation shape (60, 216, 2).

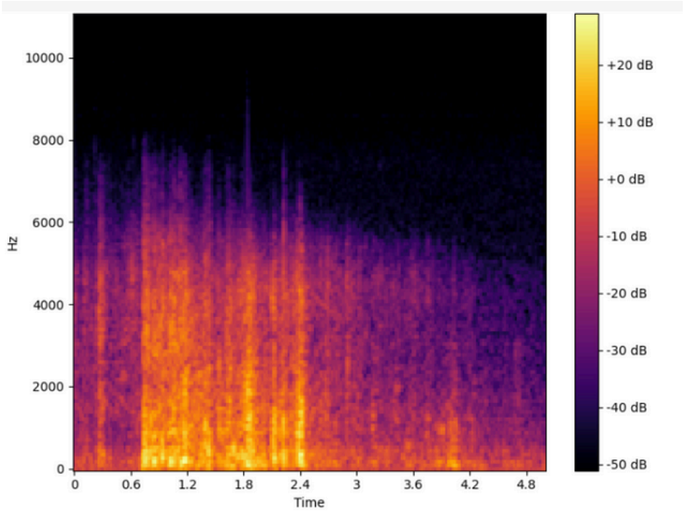


Features extraction

Università degli Studi di Padova

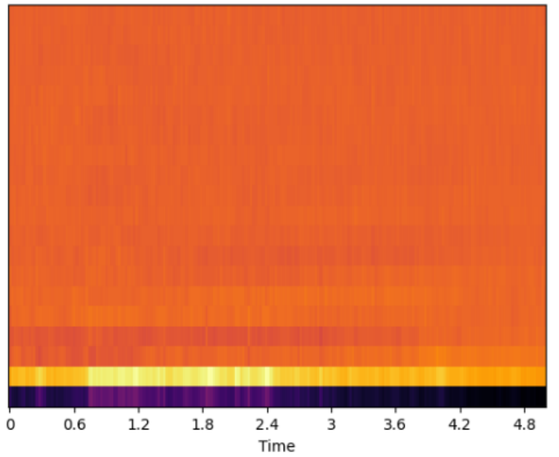
Log-Mel Spectrograms

- Window size: 1024, Hop length: 512, Mel bands: 60
- Delta features (rate of change over time) as second channel
- Input shape representation (60, 216, 2).



MFFCs

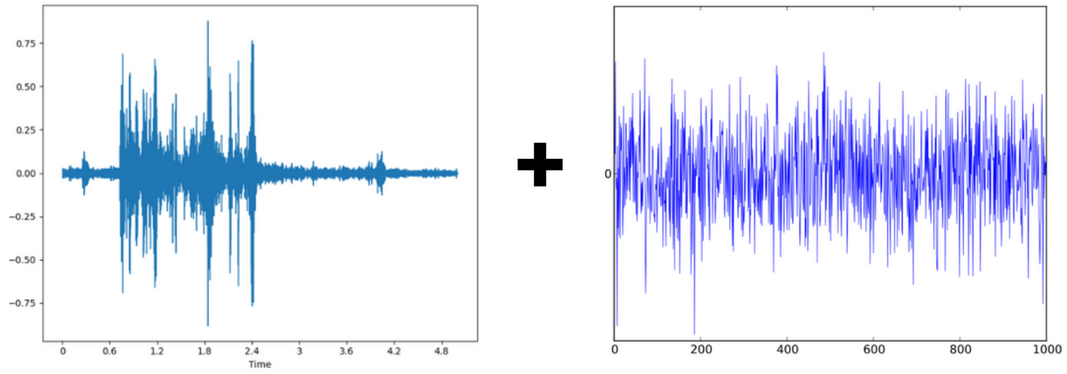
- first 20 coefficients + energy to capture the intensity over time
- Δ and Δ-Δ features used as 2 and 3 channels to capture the rate of change and acceleration
- Input shape representation (21,216,3)



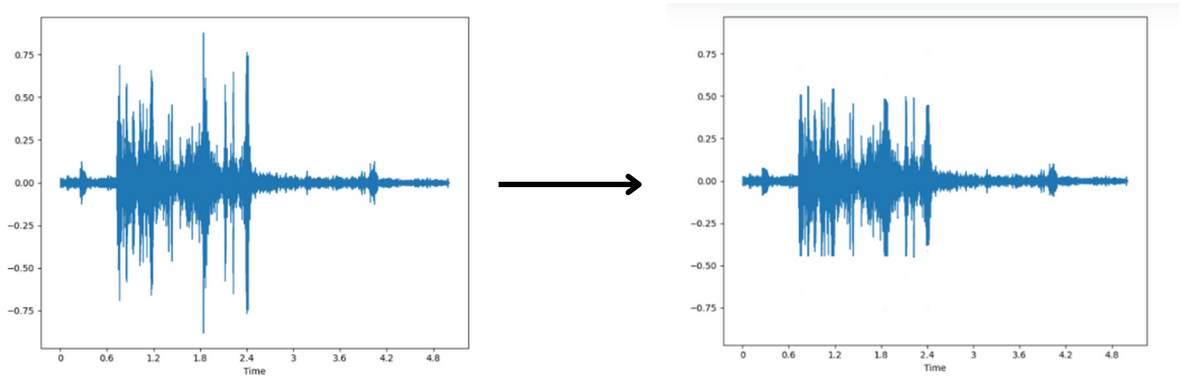
Data augmentation

To enhance the model's generalization ability and mitigate overfitting, various data augmentation techniques are employed:

White noise injection:
Adding Gaussian random noise scaled down by a factor of 0.005



Pitch shifting:
Pitch of the audio adjusted up or down while preserving its duration. Semitones considered: {-2, -1, 1, 2}.



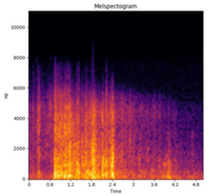
Data augmentation: Mixup

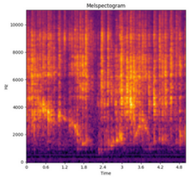
UNIVERSITÀ DEGLI STUDI DI PADOVA

Based on the work of Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification", 2018

A new training pair of features and labels (x,y) is created by combining two randomly selected feature-label pairs from the original training set

Thunderstorm log-mel spectrogram





Brushing-teeth log-mel spectrogram

Mixing factor drawn from a Beta(α,α) distribution, with α hyperparameter, e.g λ = 0.8

$$\begin{cases} x = \lambda x_i + (1-\lambda)x_j \\ y = \lambda y_i + (1-\lambda)y_j \end{cases}$$

[0.8,0.2,0,....,0]

One-hot encoded vector of the new mixing sample

[1,0,0,....,0]

One-hot encoded vector for thurderstrom class

[0,1,0,....,0]

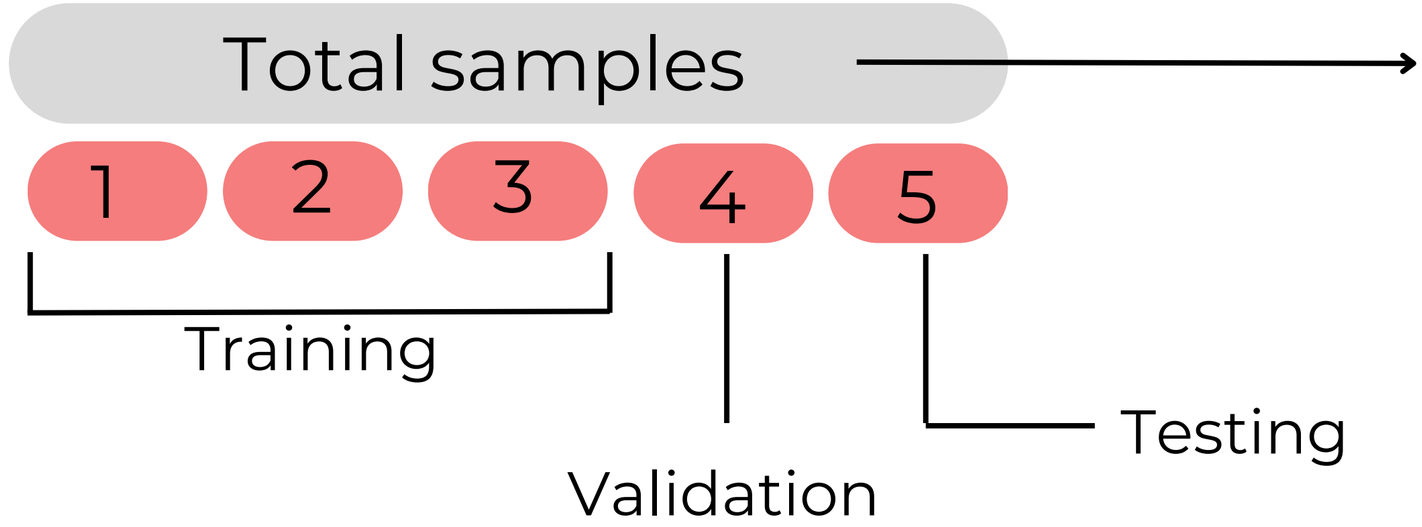
One-hot encoded vector for brushing-teeth class

Mixup applied **only** on log-mel spectrograms and MFCCs.
In case of raw audios -> distorted unrealistic artifacts

5-fold cross validation

To prevent data leakage and have a fair estimation of the metrics of the models, a 5-fold cross validation is performed using the original fold division

At each iteration:



Dataset size

Original dataset
2000 samples

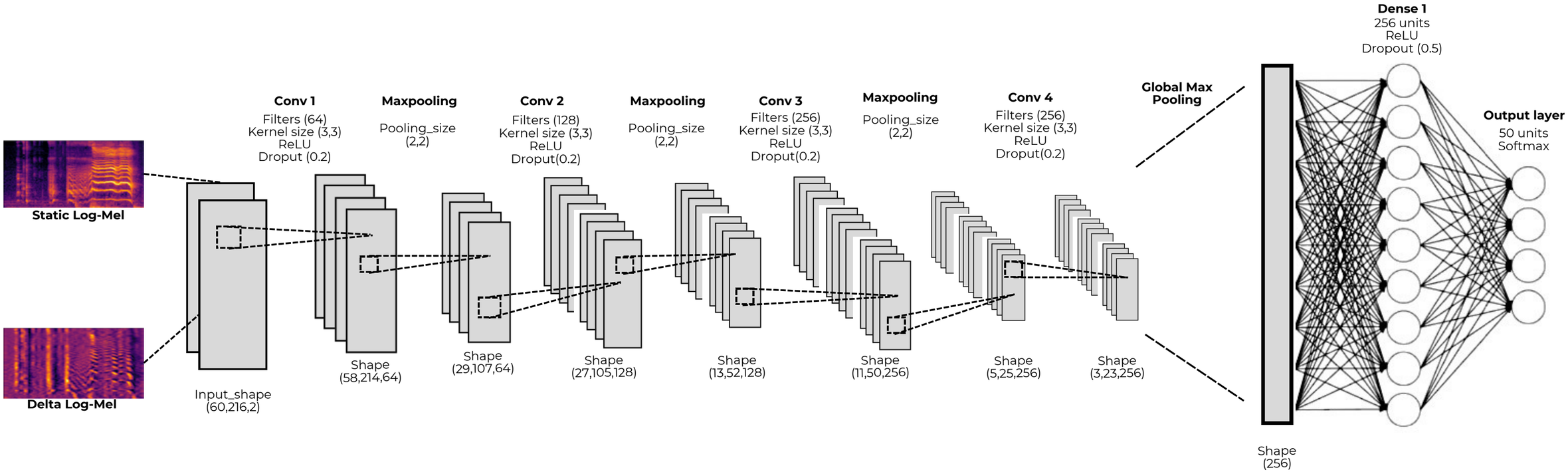
Augmentation

Augmented dataset
10000 samples
(9200 training, 400 validation, 400 testing)

Learning frameworks

Learning frameworks

1. **CNN applied to Log-Mel Spectrograms**



Training configuration: 50 epochs, batch size of 50, Adam optimizer with learning rate equal to 0.001

Learning frameworks

Università degli Studi di Padova

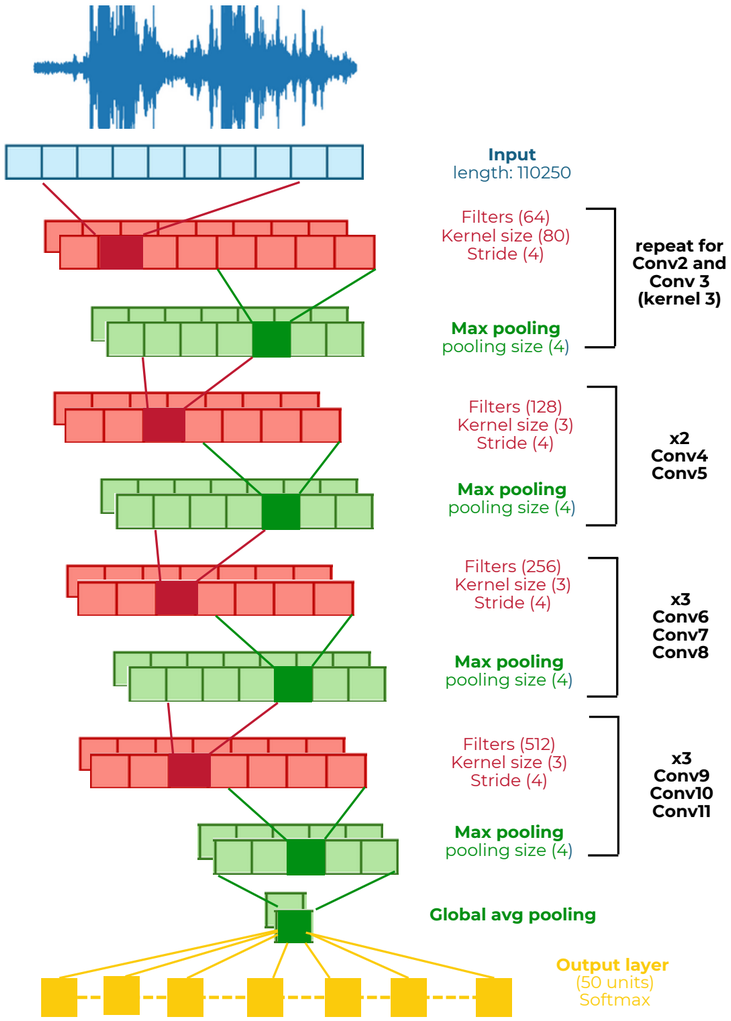
2. **CNN m11 applied to raw waveforms**

- Inspired by the work of Dai et al. “Very deep convolutional neural networks for raw waveforms”, 2017

- Fully convolutional, end-to-end stacked CNN effectively removing the need for manual feature extraction

- Relying exclusively on 1D convolutions

Training configuration: 50 epochs, batch size of 32, Adam optimizer (initial_lr = 0.001) and *ReduceLROnPlateau* strategy on the validation accuracy



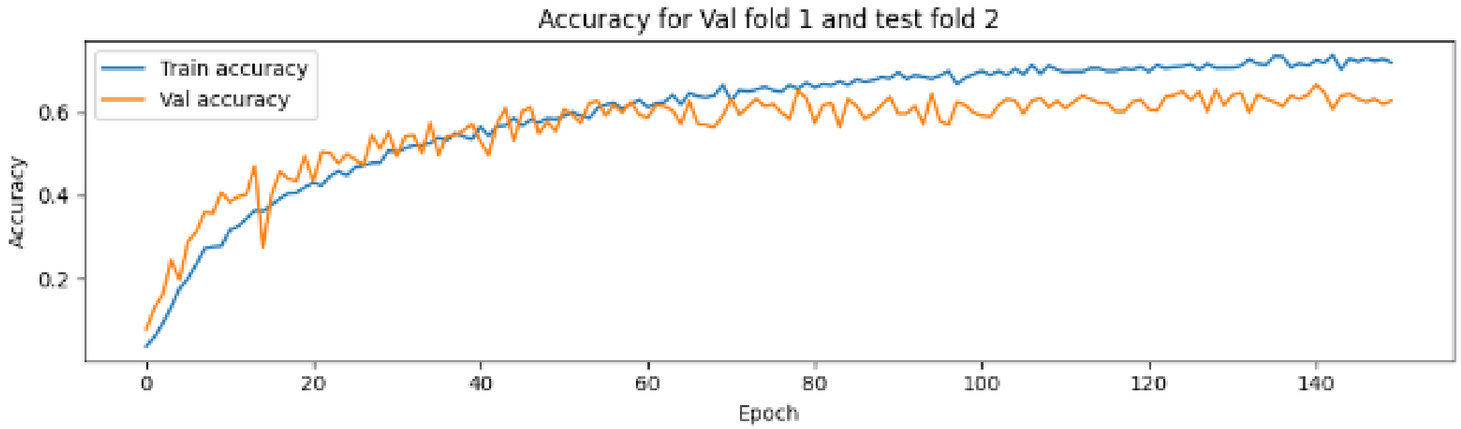
Learning frameworks

3. **CNN applied to MFCCs**

- Shallow 4-convolutional layer network similar to the one used for analyzing log-mel spectrograms but with 32,64,128 and 256 filters

- Different regularisation configuration

Training configuration: 150 epochs, batch size of 50, Adam optimizer with learning rate equal to 0.001



Learning frameworks

4. **CRNN applied to log-mel spectrograms**

This approach combines the strengths of CNNs and RNNs since it can identify both spatial and temporal patterns in the audios

Log-Mel CNN

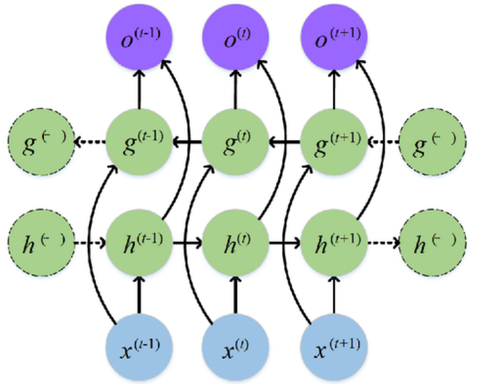
initial feature extractor

→

max pooling (axis =2, time dimension)

aggregate temporal information and preserve critical frequency components

→



2 Bidirectional Gated Recurrent Units (GRU) with 256 units

→

Output layer with 50 units and softmax

Training configuration: 150 epochs, batch size of 50, Adam optimizer (initial_lr = 0.001) and *ReduceLROnPlateau* strategy on the validation accuracy

Learning frameworks

Ensemble models

Purpose: Combine predictions from multiple models to enhance accuracy and robustness.

Ensemble Techniques:

Product of Probabilities:

- Amplifies high-confidence predictions common across models.
- **Strength**: Focuses on areas of agreement.

Average of Probabilities (Soft Voting):

- Smooths predictions by giving equal weight to each model's output.
- **Strength**: Reduces bias from any single model

Results

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Results

UNIVERSITÀ DEGLI STUDI DI PADOVA

- All experiments are conducted in the Kaggle environment, utilizing a P100 GPU with 32 GB memory

kaggle

- Evaluation based on the following metrics: accuracy, precision, F1 score, and training time. We do not need any adjustments for class imbalance.

Model	Parameters	Memory (Mb)
LogMelCNN	1,041,778	3.97
RawCNN	1,811,442	6.91
MFCCNet	468,978	1.79
LogMelRCNN	2,961,010	11.30

TABLE 1: Parameters and memory from each architecture.

Results: Data augmentation

Augmentation on raw waveforms

Performances

Why?

- White noise and pitch shift directly affect the waveforms,
- This confuse the model
- No decomposition of the signal into higher-level features.

Augmentation	Accuracy	Training time (s)
No aug	60.9 ± 0.03	209
White noise	60.2 ± 0.03	940
Pitch shifting	60.6 ± 0.02	931

TABLE 2: Results of Raw-CNN for different types of augmentation techniques

Results: Data augmentation

Università degli Studi di Padova

Augmentation on log-mel spectrograms and MFCCs

Augmentation	Accuracy	Training time (s)
No aug	63.8 ± 0.02	47
White noise	67.4 ± 0.04	173
Pitch shifting	68.9 ± 0.02	175
Mixup	71.4 ± 0.02	244
All	69.3 ± 0.03	292

TABLE 3: Results of Logmel-CNN for different types of augmentation techniques

Augmentation	Accuracy	Training time (s)
No aug	58.0 ± 0.03	50
White noise	60.8 ± 0.05	204
Pitch shifting	62.9 ± 0.03	173
Mixup	65.7 ± 0.04	249
All	63.6 ± 0.04	246

TABLE 4: Results of MFCC-CNN for different types of augmentation techniques

Performances

Training times

Results: Data augmentation

Università degli Studi di Padova

Augmentation on log-mel spectrograms and MFCCs

Augmentation	Accuracy	Training time (s)
No aug	63.8 ± 0.02	47
White noise	67.4 ± 0.04	173
Pitch shifting	68.9 ± 0.02	175
Mixup	71.4 ± 0.02	244
All	69.3 ± 0.03	292

TABLE 3: Results of Logmel-CNN for different types of augmentation techniques

Augmentation	Accuracy	Training time (s)
No aug	58.0 ± 0.03	50
White noise	60.8 ± 0.05	204
Pitch shifting	62.9 ± 0.03	173
Mixup	65.7 ± 0.04	249
All	63.6 ± 0.04	246

TABLE 4: Results of MFCC-CNN for different types of augmentation techniques

Pitch shifting has the highest single-augmentation accuracy: it adds valuable variation to the data, more effectively than white noise

Results: Data augmentation

Università degli Studi di Padova

Augmentation on log-mel spectrograms and MFCCs

Augmentation	Accuracy	Training time (s)
No aug	63.8 ± 0.02	47
White noise	67.4 ± 0.04	173
Pitch shifting	68.9 ± 0.02	175
Mixup	71.4 ± 0.02	244
All	69.3 ± 0.03	292

TABLE 3: Results of Logmel-CNN for different types of augmentation techniques

Augmentation	Accuracy	Training time (s)
No aug	58.0 ± 0.03	50
White noise	60.8 ± 0.05	204
Pitch shifting	62.9 ± 0.03	173
Mixup	65.7 ± 0.04	249
All	63.6 ± 0.04	246

TABLE 4: Results of MFCC-CNN for different types of augmentation techniques

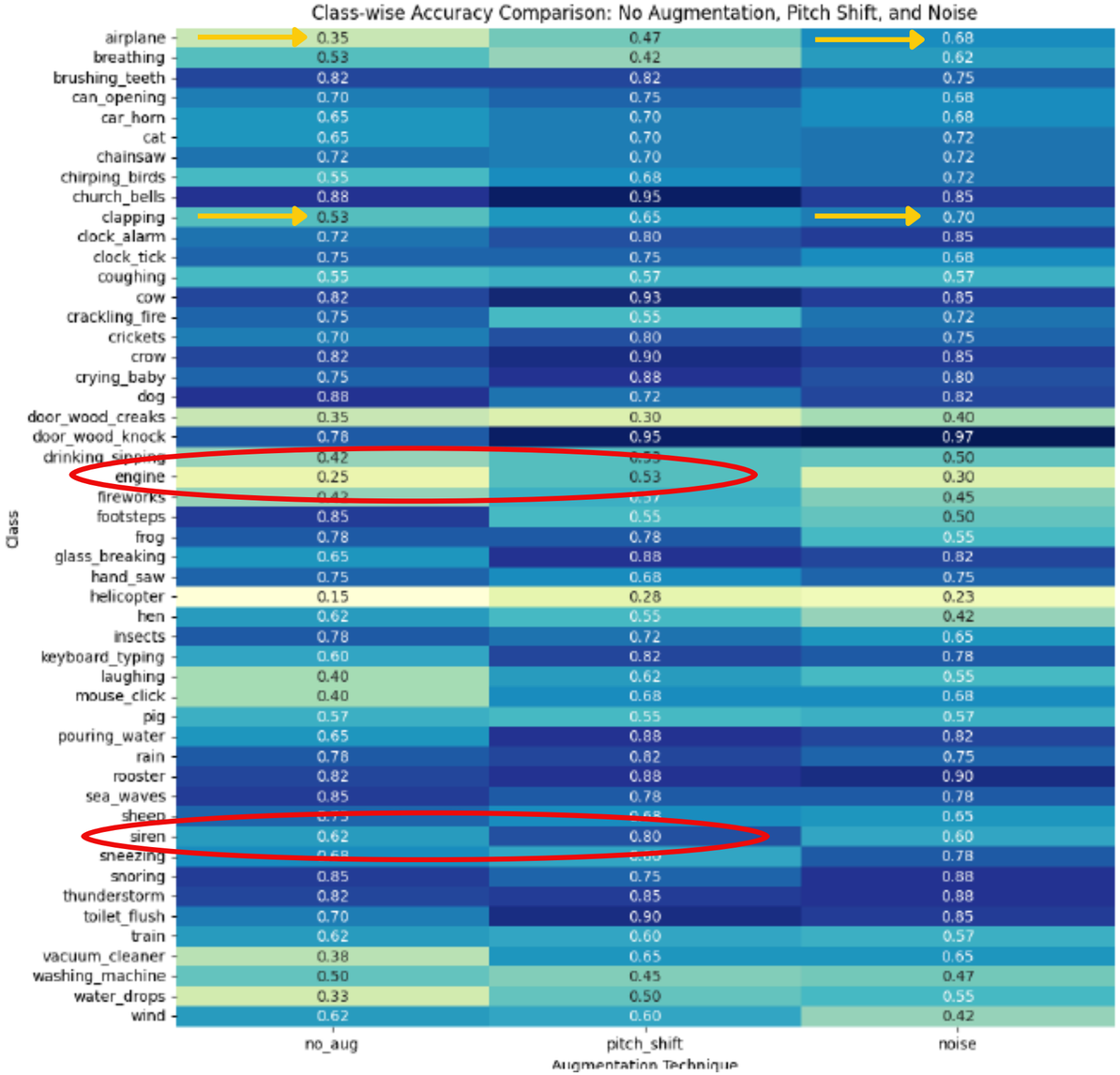
Mix up **alone** reaches the highest score. Why?

If applied to already augmented data, mixup may produce unnatural hybrid samples, potentially confusing the model

Results: Data augmentation

Class-wise accuracy comparison

- **Pitch augmentation**: beneficial for classes with a **dominant frequency** : *church bells*, *siren*, and *engine* noises.
- In the real world, their pitch vary **depending on their source**
- **Noise augmentation** helpful for sounds that naturally occur in environments with background noises, like *airplane*, *clapping*, *sneezing* or *breathing*



Comparison of the models

Università degli Studi di Padova

Model	Accuracy	Precision	F1-score	Training time (s)
LogMelCNN	71.40 ± 0.02	75.05 ± 0.03	70.50 ± 0.03	244
RawCNN	60.95 ± 0.03	62.84 ± 0.04	59.79 ± 0.03	209
MFCCNet	65.70 ± 0.04	71.61 ± 0.03	64.88 ± 0.04	249
LogMelCRNN	74.80 ± 0.03	76.99 ± 0.03	73.96 ± 0.03	2102
Ensemble (Average)	78.65 ± 0.02	80.67 ± 0.02	77.67 ± 0.03	2804
Ensemble (Product)	79.10 ± 0.03	81.13 ± 0.03	78.25 ± 0.03	2804

Among the **individual** models:

- The **LogMelCRNN** performs the **best**
- The **RawCNN** performs **the worst**

Comparison of the models

UNIVERSITÀ DEGLI STUDI DI PADOVA

Model	Accuracy	Precision	F1-score	Training time (s)
LogMelCNN	71.40 ± 0.02	75.05 ± 0.03	70.50 ± 0.03	244
RawCNN	60.95 ± 0.03	62.84 ± 0.04	59.79 ± 0.03	209
MFCCNet	65.70 ± 0.04	71.61 ± 0.03	64.88 ± 0.04	249
LogMelCRNN	74.80 ± 0.03	76.99 ± 0.03	73.96 ± 0.03	2102
Ensemble (Average)	78.65 ± 0.02	80.67 ± 0.02	77.67 ± 0.03	2804
Ensemble (Product)	79.10 ± 0.03	81.13 ± 0.03	78.25 ± 0.03	2804

Difference between LogMelCNN and MFCCNet:

- Log-mel spectrograms retain detailed frequency information over time
- MFCCs compress spectral information, leading to loss of finer details, critical for distinguish sounds

Comparison of the models

UNIVERSITÀ DEGLI STUDI DI PADOVA

Model	Accuracy	Precision	F1-score	Training time (s)
LogMelCNN	71.40 ± 0.02	75.05 ± 0.03	70.50 ± 0.03	244
RawCNN	60.95 ± 0.03	62.84 ± 0.04	59.79 ± 0.03	209
MFCCNet	65.70 ± 0.04	71.61 ± 0.03	64.88 ± 0.04	249
LogMelCRNN	74.80 ± 0.03	76.99 ± 0.03	73.96 ± 0.03	2102
Ensemble (Average)	78.65 ± 0.02	80.67 ± 0.02	77.67 ± 0.03	2804
Ensemble (Product)	79.10 ± 0.03	81.13 ± 0.03	78.25 ± 0.03	2804

- The **ensemble models** provide the **best** performances: accuracy comparable to the **average human** classification of **81.3%**
- The trade-off between training time and performance: CRNN and ensemble models have **longer training times**

Conclusions

In summary:

- Models that used **compressed representations**, such as MFCCs, or raw waveforms performed **less effectively**
- It's important to preserve **detailed frequency information** in this classification task
- **Combining models** with different input features and architectures **enhances** the system's ability to correctly classify sounds
- Model performance could be further improved by using **class-specific data augmentation strategies**

Let's try one of our models