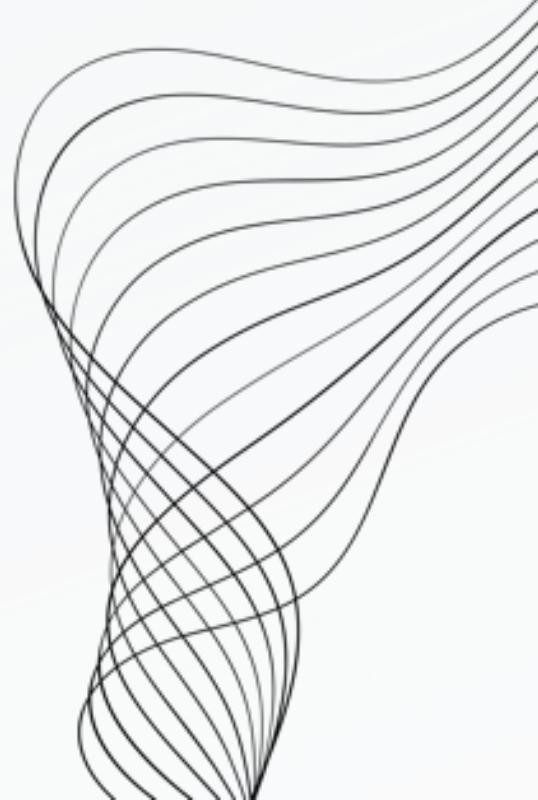


# **DEEP LEARNING MODELS FOR ENVIRONMENTAL SOUND CLASSIFICATION: A COMPARATIVE APPROACH**

*Project B2. Audio Classification Task.*



# CONTENT

01

OUR GOAL

02

DATASET DESCRIPTION

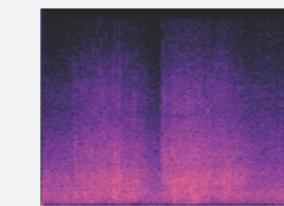


03

DATA PREPROCESSING

04

AUDIO REPRESENTATIONS

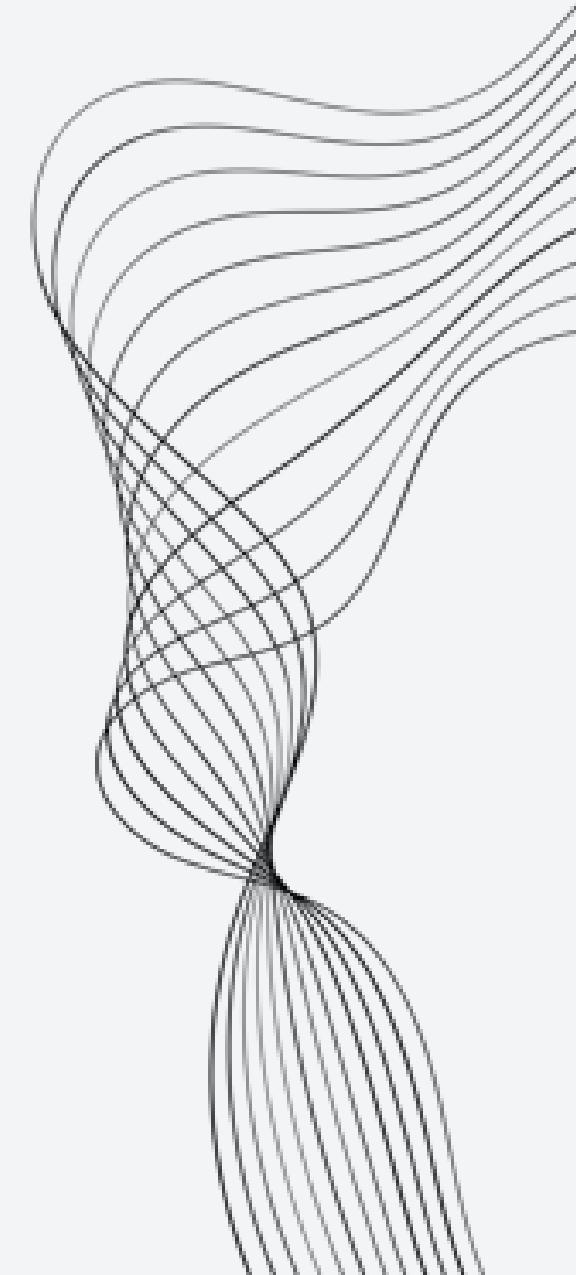
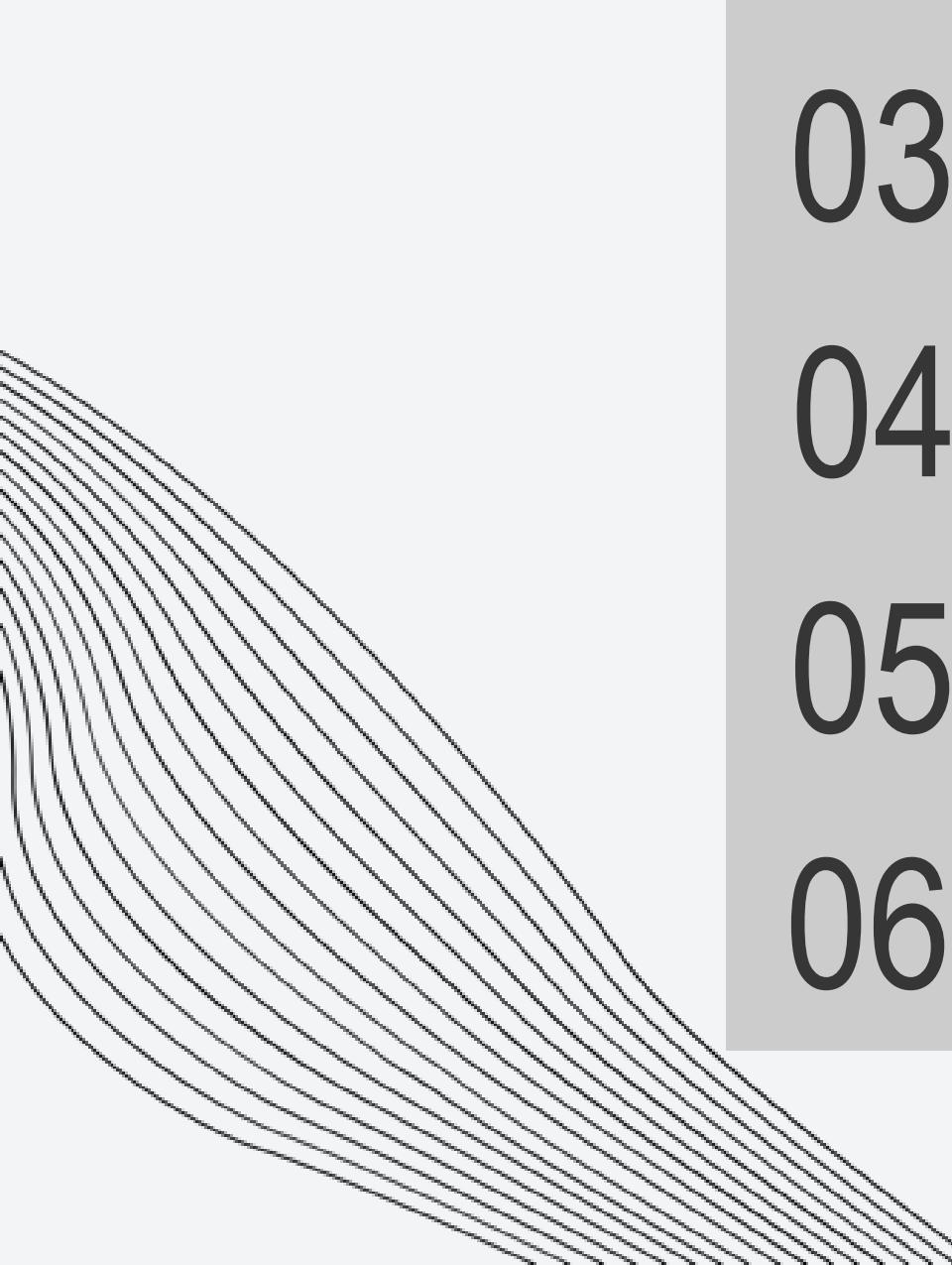


05

MODEL ARCHITECTURES

06

RESULTS AND COMPARISONS





# GOAL

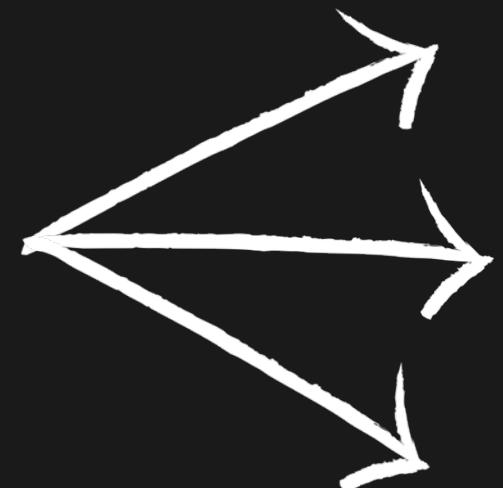
## AUDIO CLASSIFICATION TASK.



SPECTs or MFCC



CNN, RNN, CRNN



PREDICTIONS

# ESC-10 DATASET:

## What is it?

A subgroup of the well-known ESC-50 dataset.

## What is it composed of?

Labeled data of environmental sounds, such as the sound of rain, rooster crowing, baby crying...

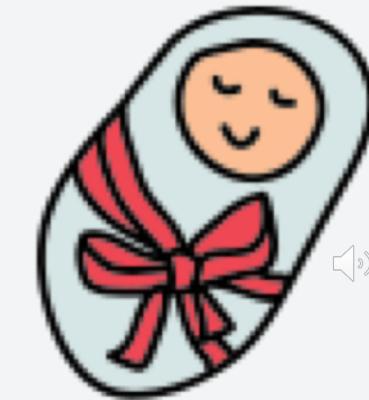
## Main characteristics:

10 classes, 400 data, 40 per each class. 5 seconds per audio. Sounds clearly audible and classifiable by the human ear.



# DATA AUGMENTATION

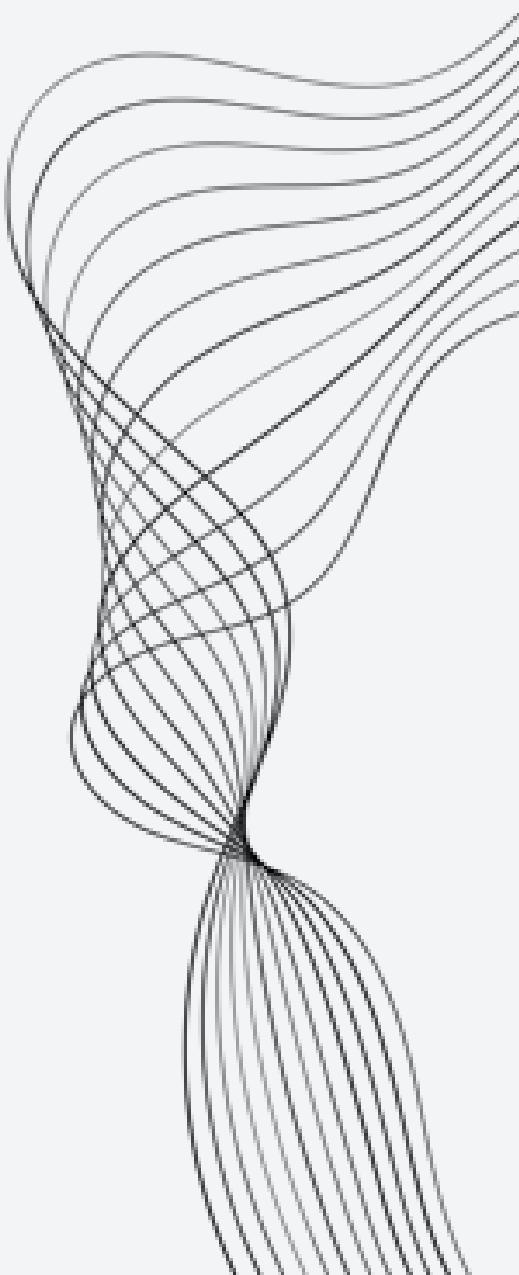
Harmonic Sound → Pitch Shifting



Soundscape → Adding Noise

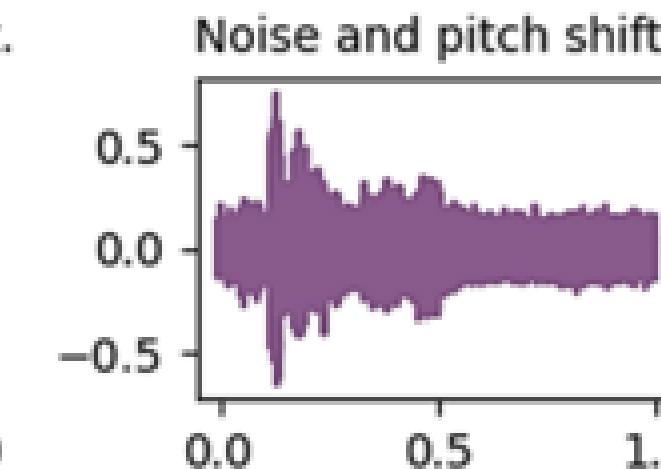
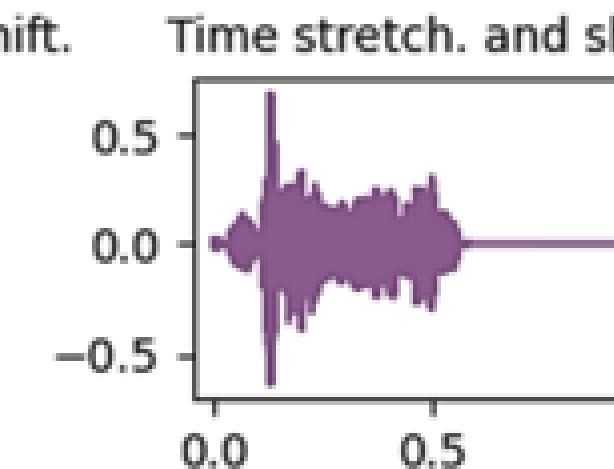
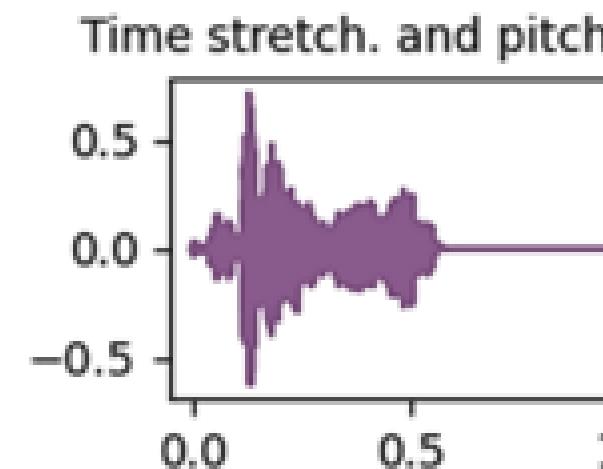
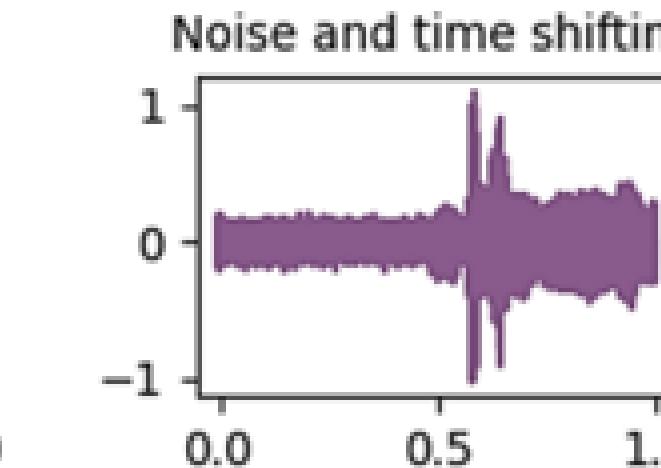
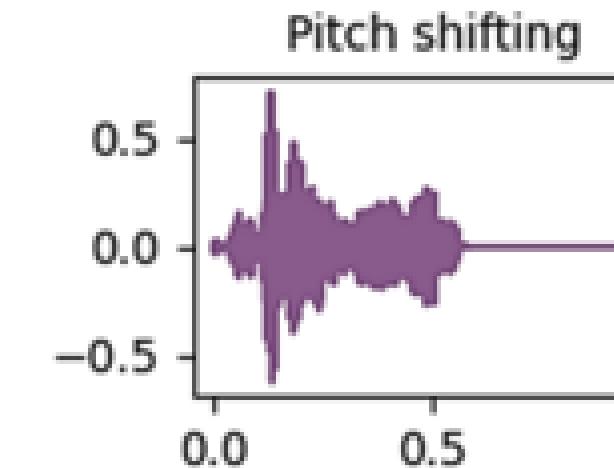
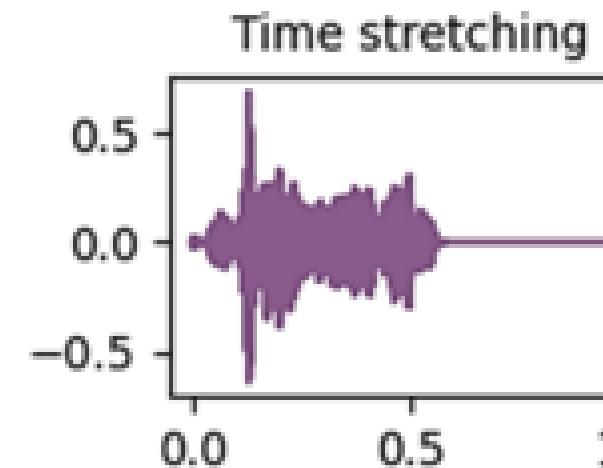
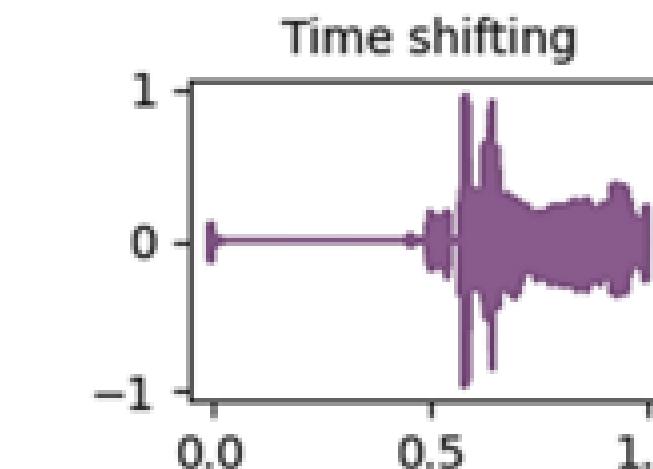
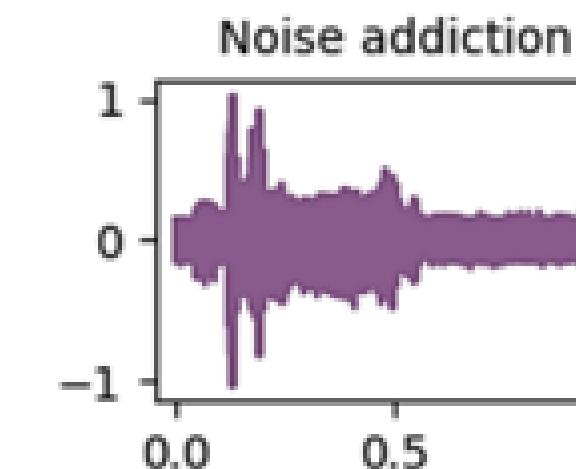
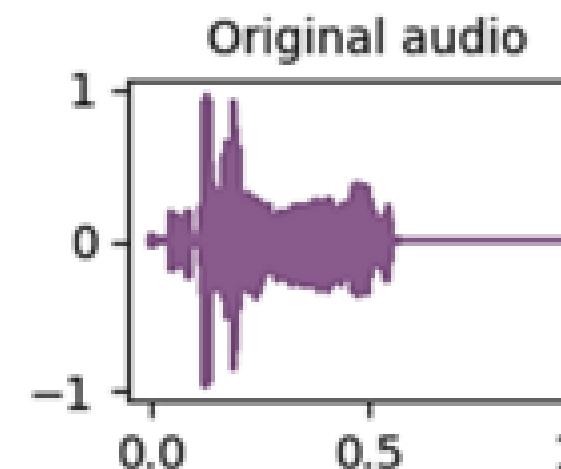


Percussive Sound → Time Shifting  
Time stretching

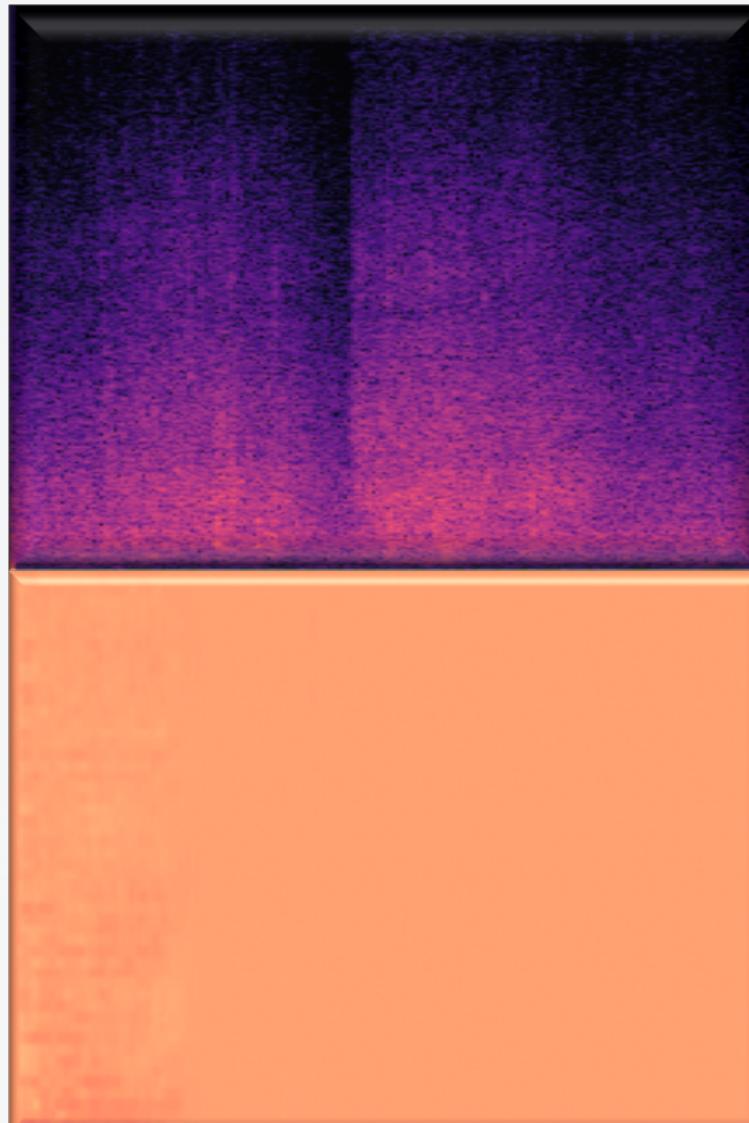


# DATA AUGMENTATION

(of a  )



# AUDIO REPRESENTATION



## SPECTROGRAM

Time VS frequency. Colors represent the magnitude or power of the frequencies at each point in time.

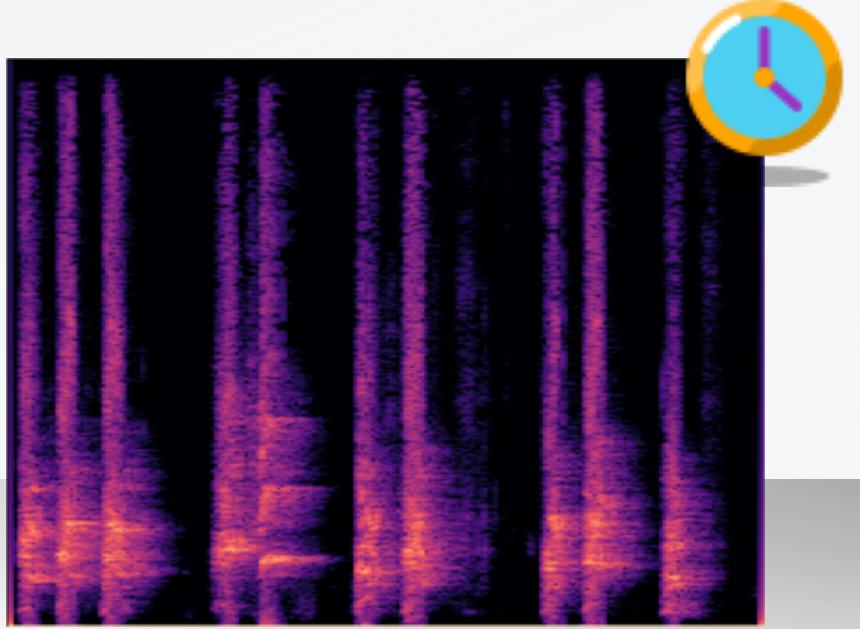
## MFCC

Feature representation. It captures the spectral characteristics of a signal.

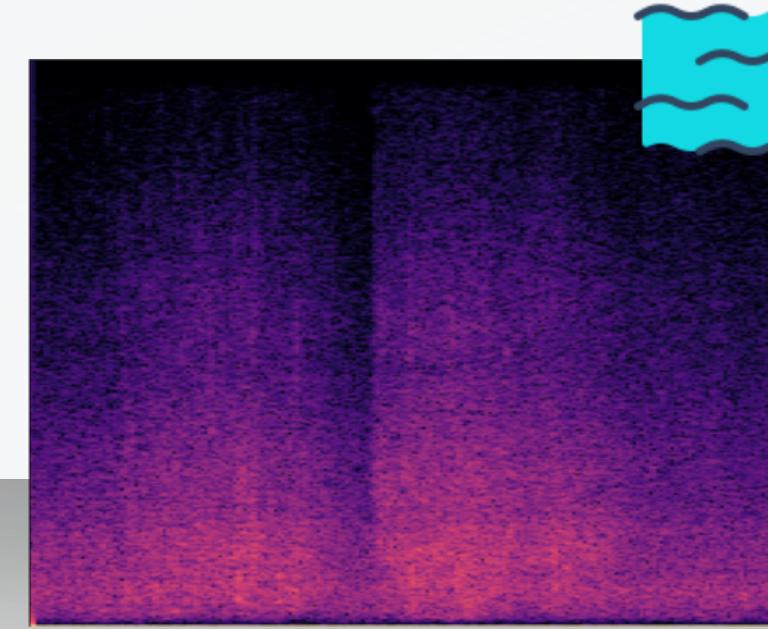
# AUDIO REPRESENTATION: SPECTROGRAM

Reasons why they might  
be a good choice:

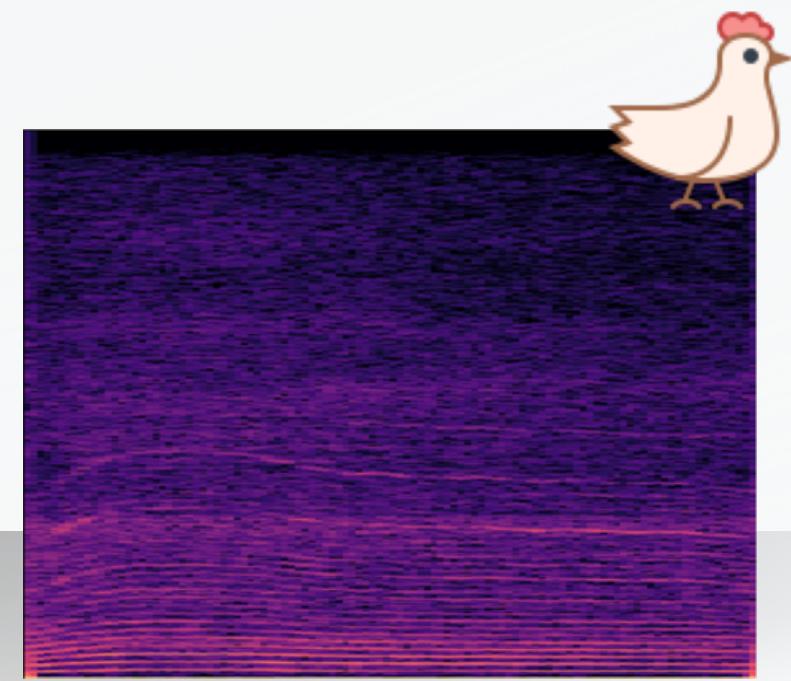
More sensitive to  
sound traits.



Percussive  
Sound



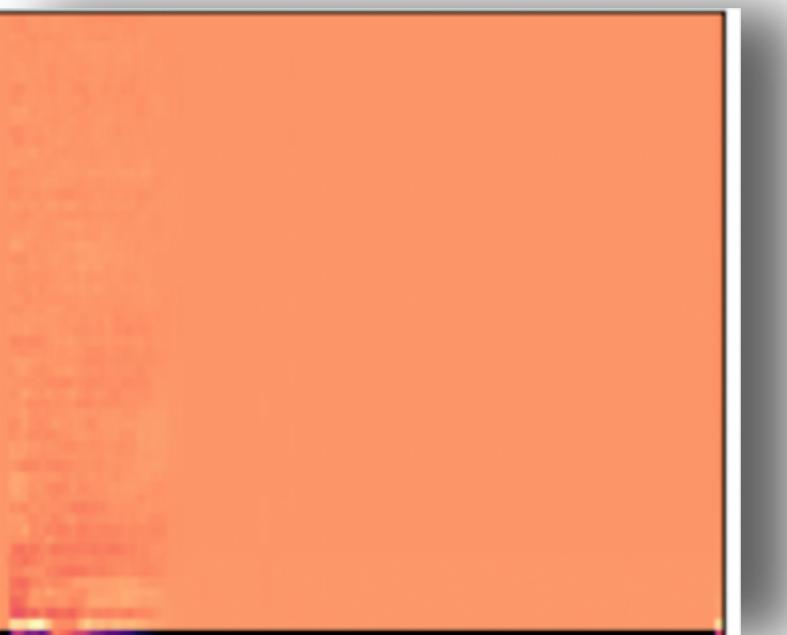
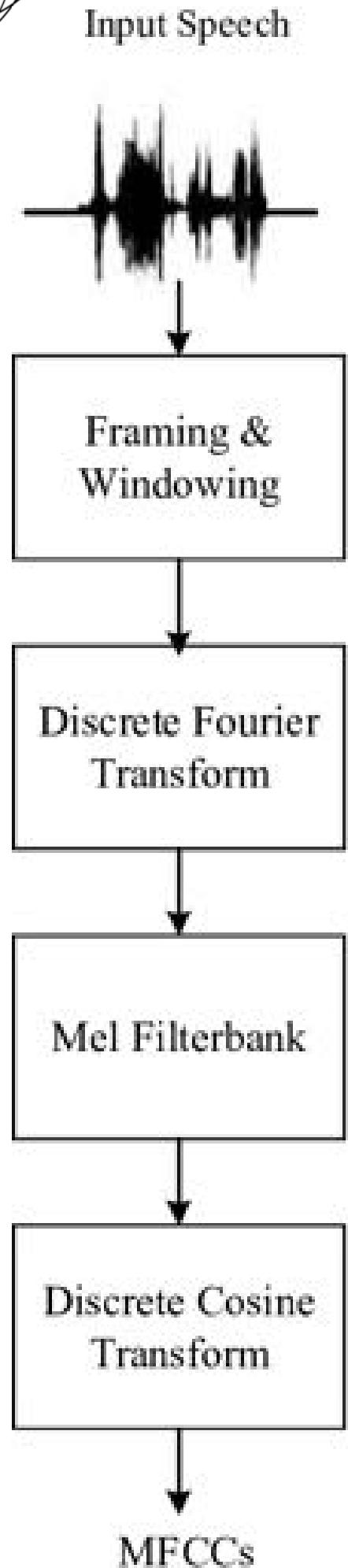
Soundscape



Harmonic Sound



# AUDIO REPRESENTATION: MFCC



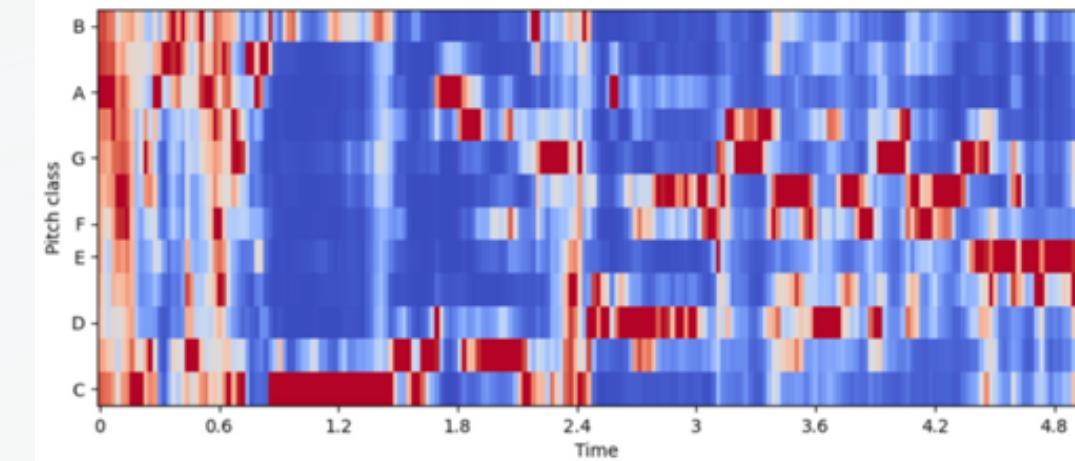
- Hop Length: 512
- Win Length: 1024
- Num. of MFCC: 60

# AUDIO REPRESENTATION: MFCC

Reasons why they might  
be a good choice:

Highly appreciated in  
the scientific literature  
of this task

Integrate chromatogram  
in input, why?



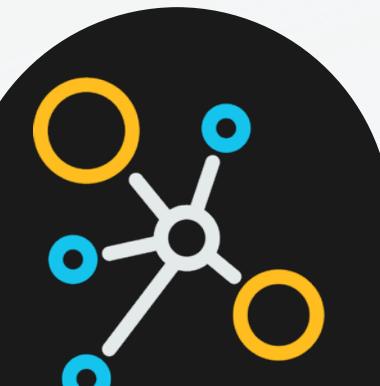
Less sensitive to noise  
and temporal information

# MODELS



NN based on convolutional filters that enable the NN to automatically learn and recognize patterns, shapes, and structures within the data.

## CNN



NN designed to handle long-term dependencies and capture sequential patterns in data.

## RNN (LSTM)



Hybrid NN architecture that combines convolutional layers for spatial feature extraction and recurrent layers for temporal modeling.

## CRNN

# CHARATERISTICS THAT ARE COMMON TO ALL MODELS IMPLEMENTED



## ACTIVATION FUNCTION

Softmax in the last layer. ReLU for the others.

## LOSS FUNCTION

Categorical Crossentropy

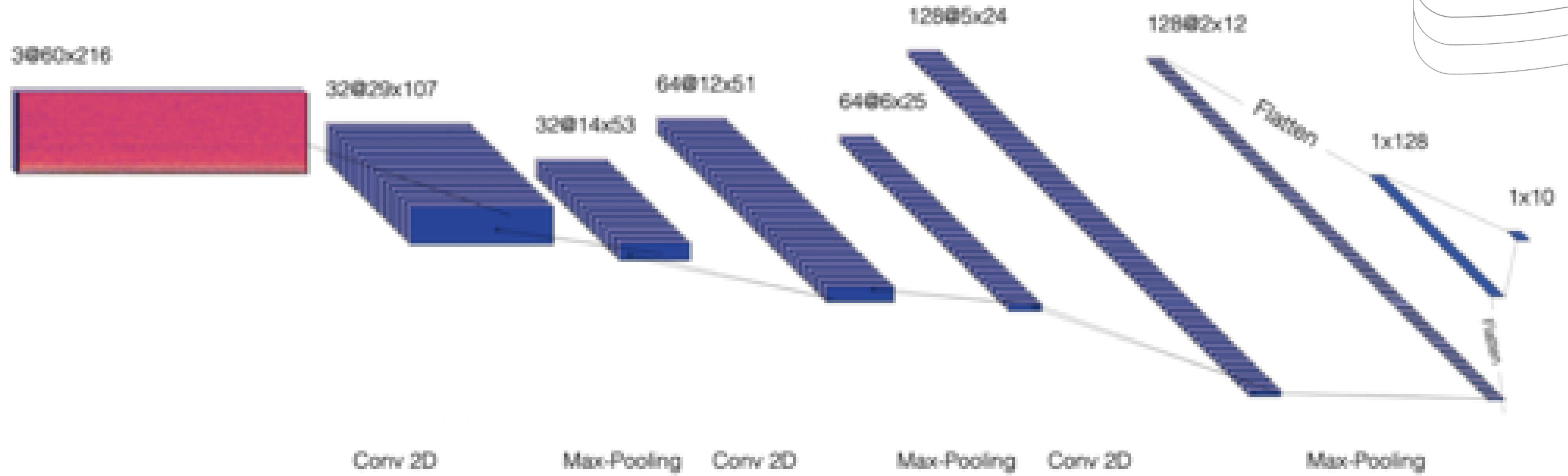
## OPTIMIZER

ADAM

## EPOCHS CRITERIA

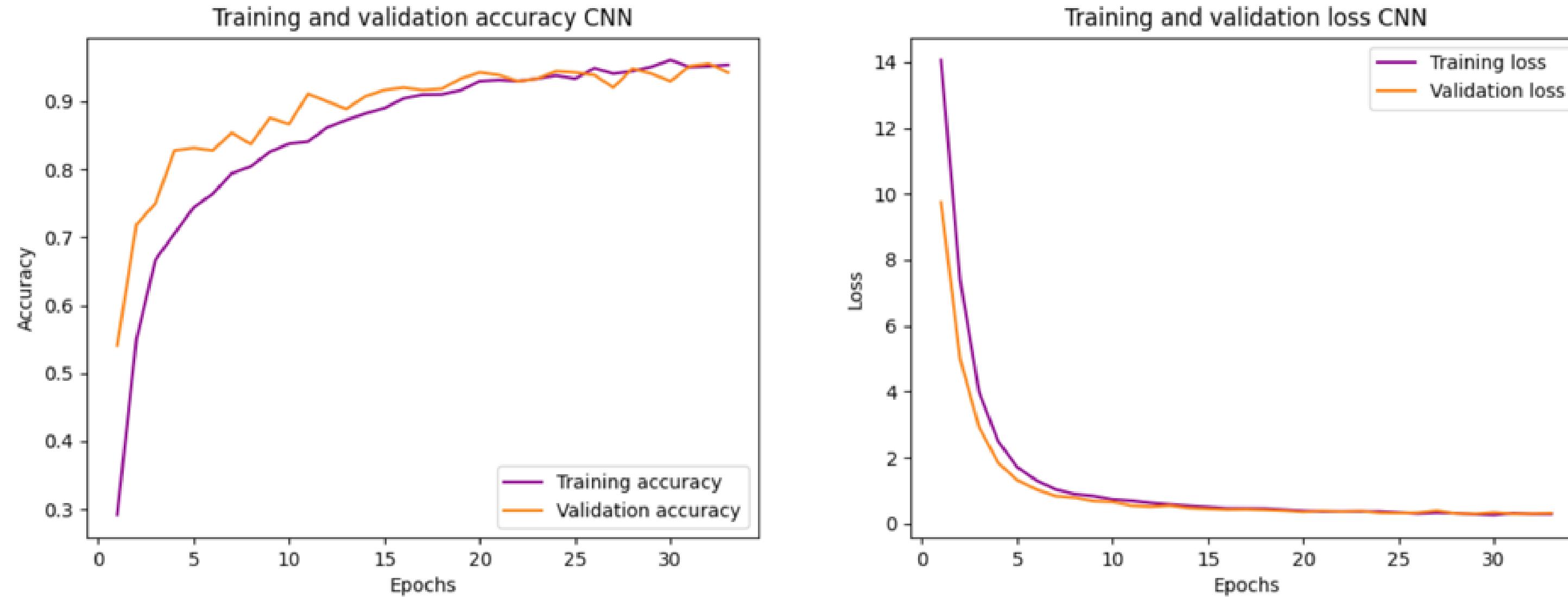
Early stopping on validation loss with patience = 4 to choose the number of epochs.

# CNN ON MFCC



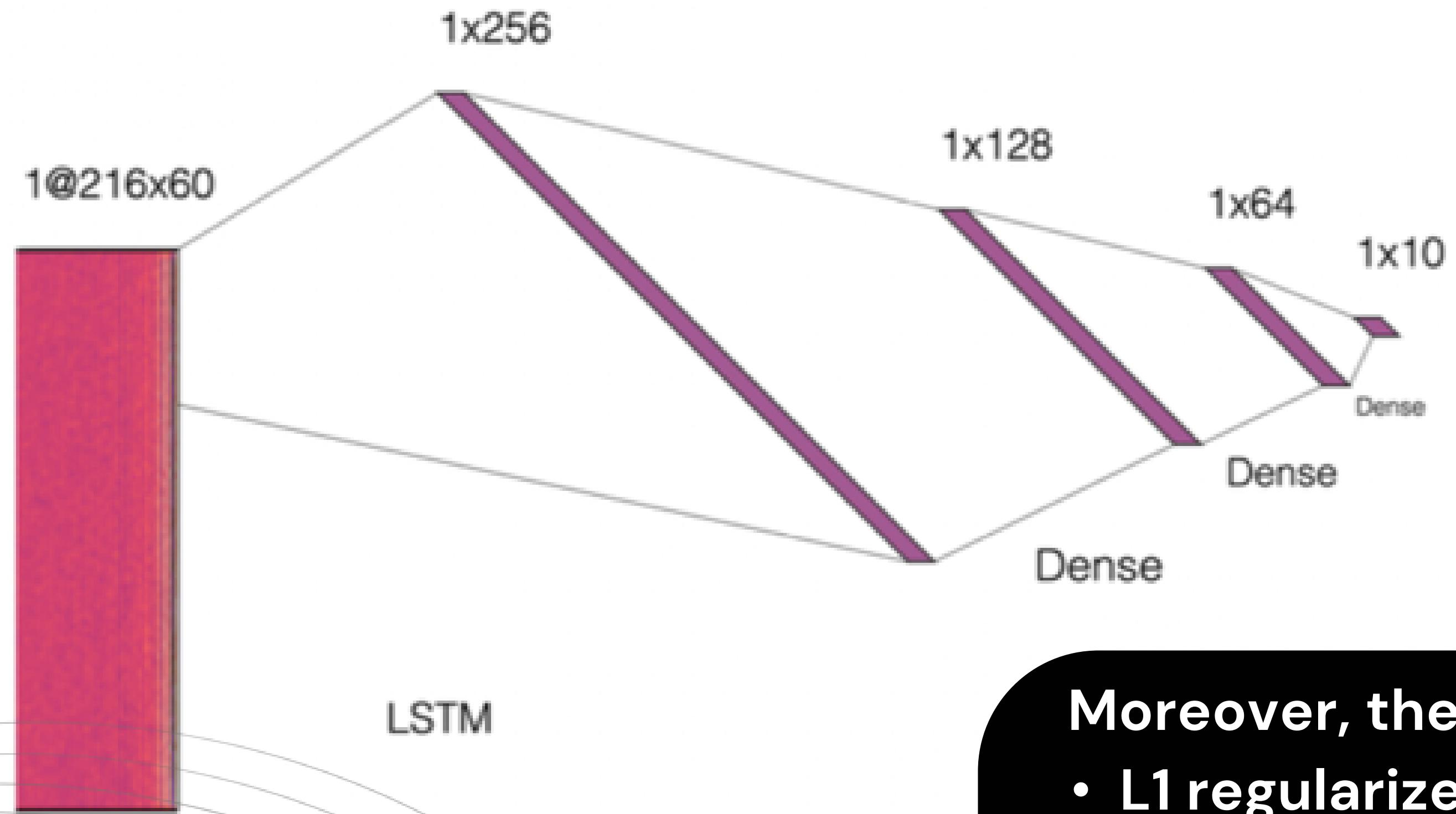
- Moreover, the use of:
- L1 regularizer (0.01)
  - Dropout (0.2)
  - 33 epochs

# PERFORMANCE



Accuracy	Size of model (KB)	Inference time (s)
94.07	1622.69	0.28

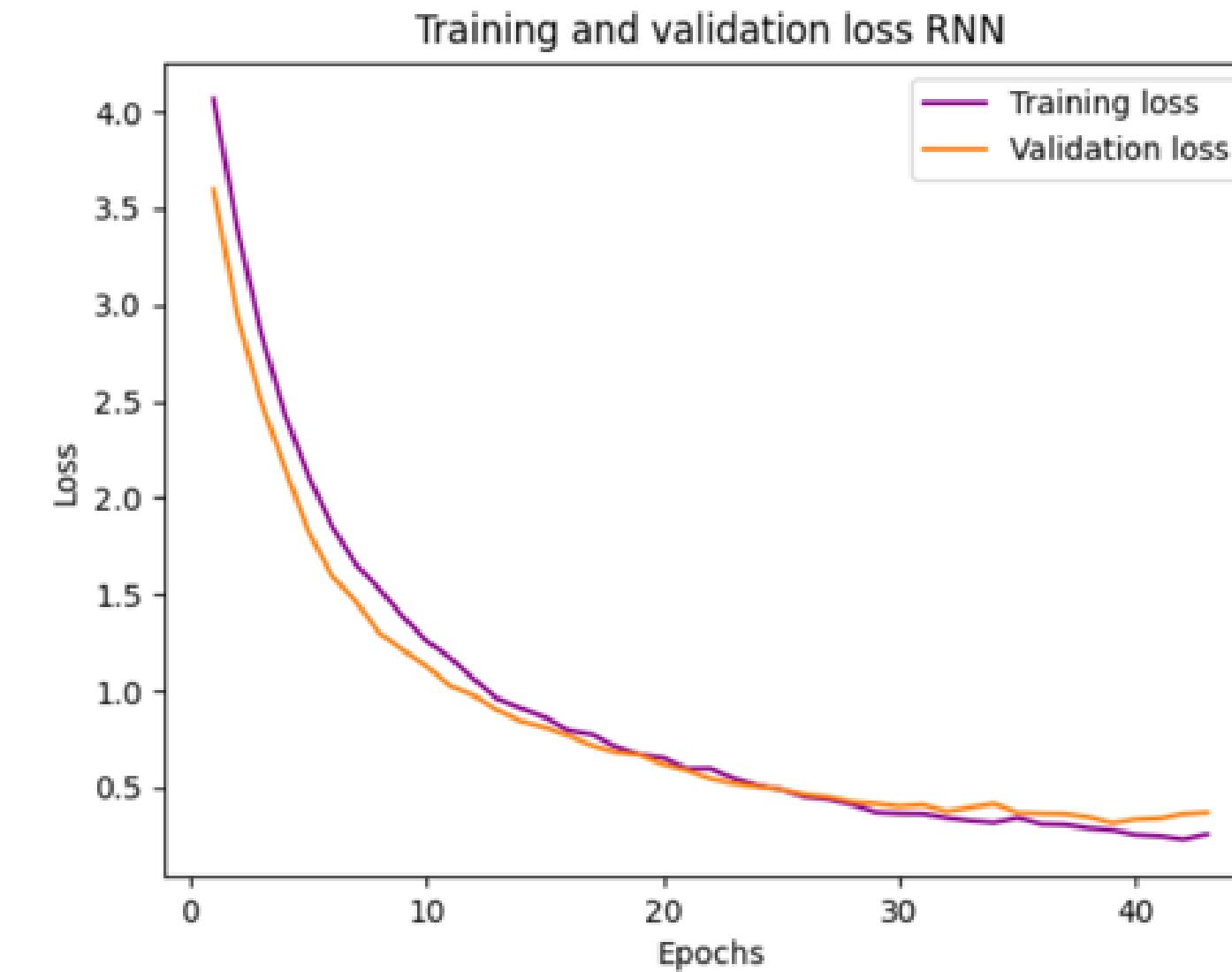
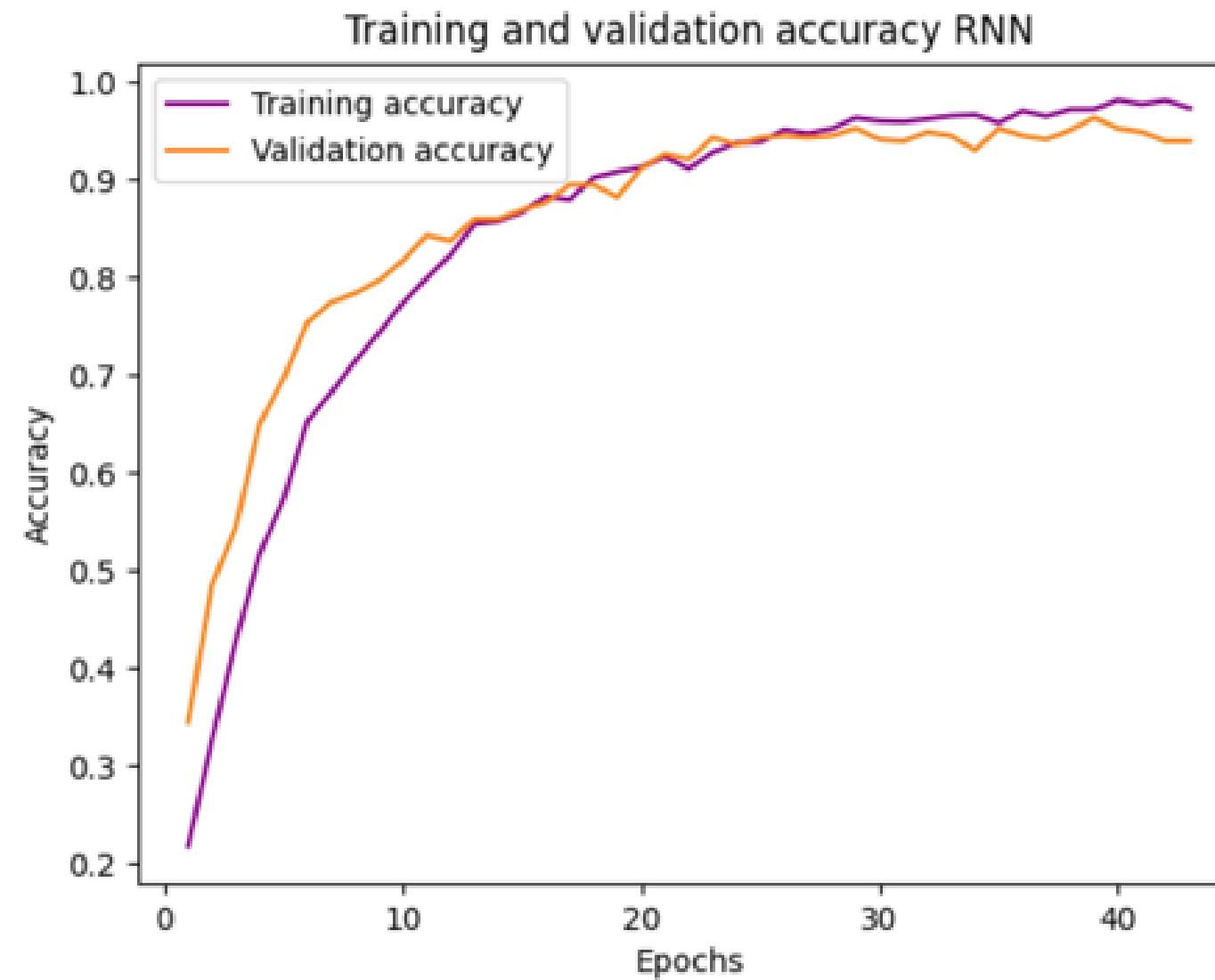
# LSTM ON MFCC



Moreover, the use of:

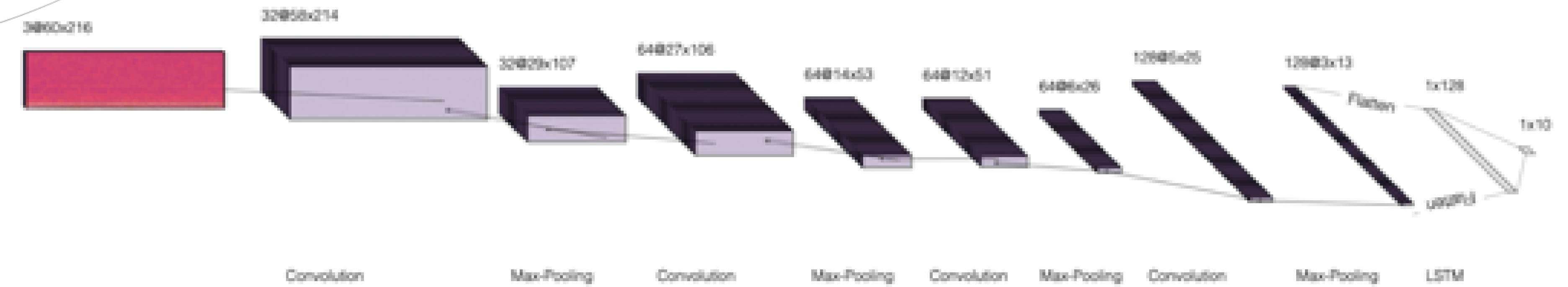
- L1 regularizer (0.001)
- Dropout (0.2)
- 43 epochs

# PERFORMANCE



Accuracy	Size of model (KB)	Inference time (s)
95.93	1334.07	0.2

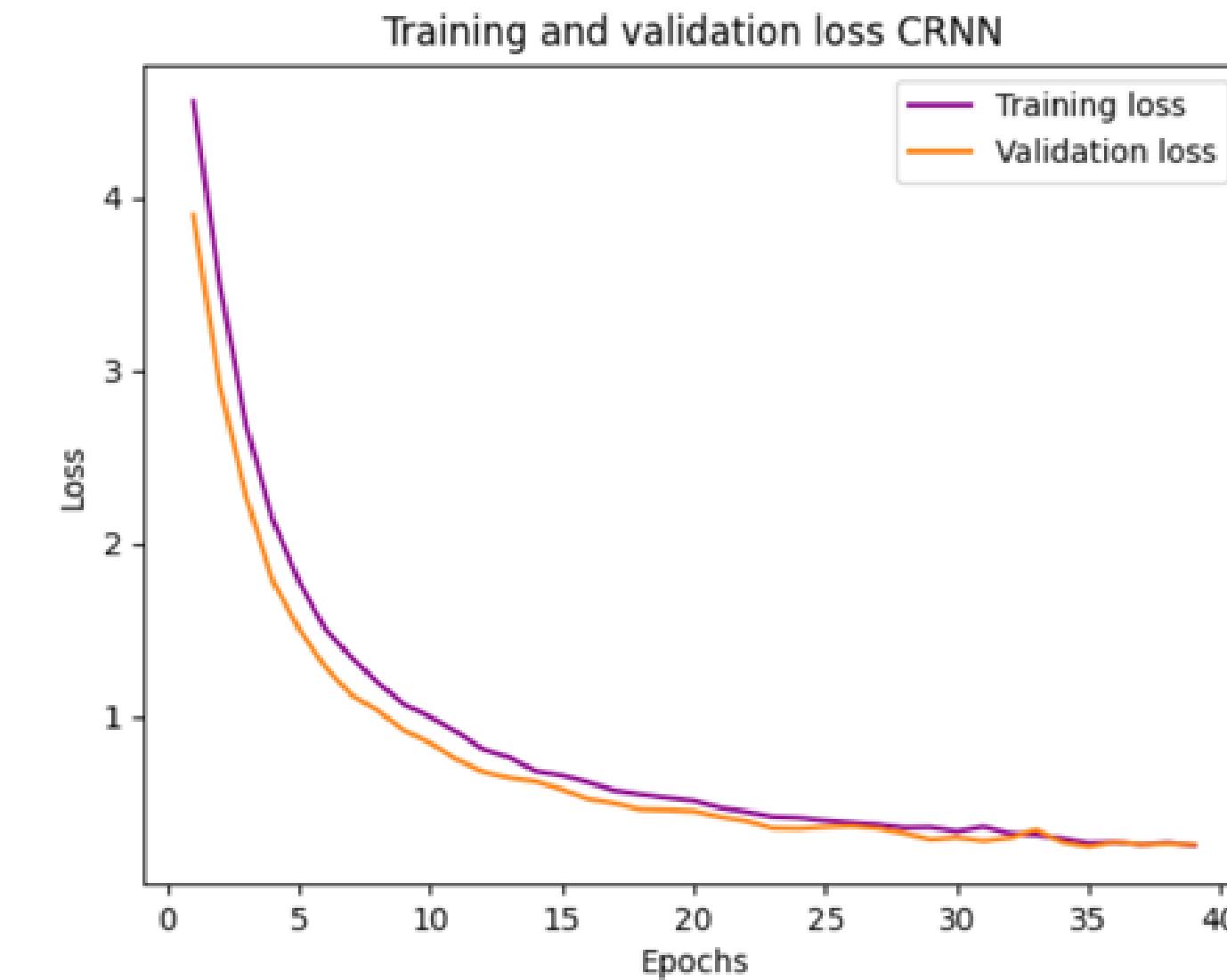
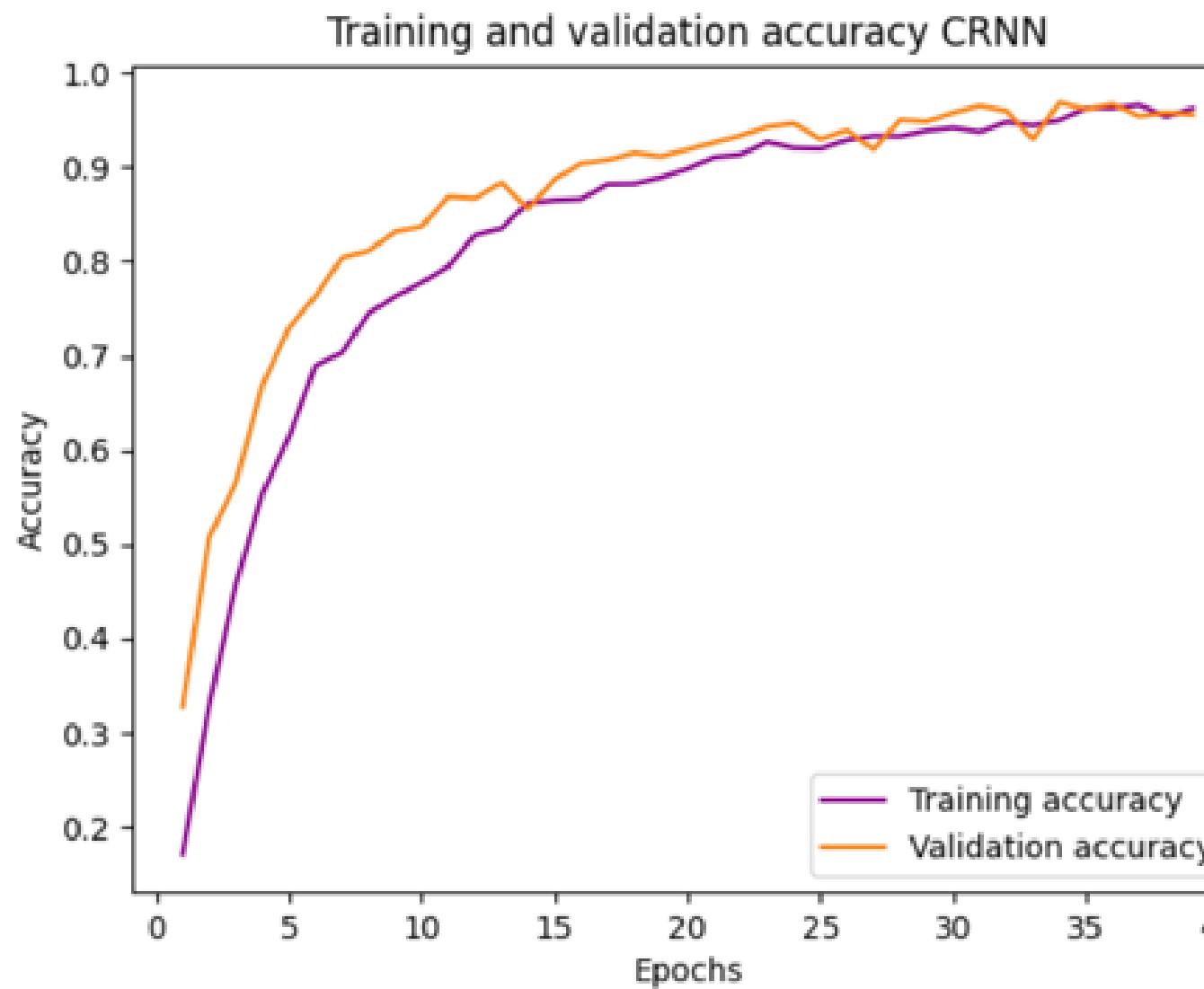
# CRNN ON MFCC



Moreover, the use of:

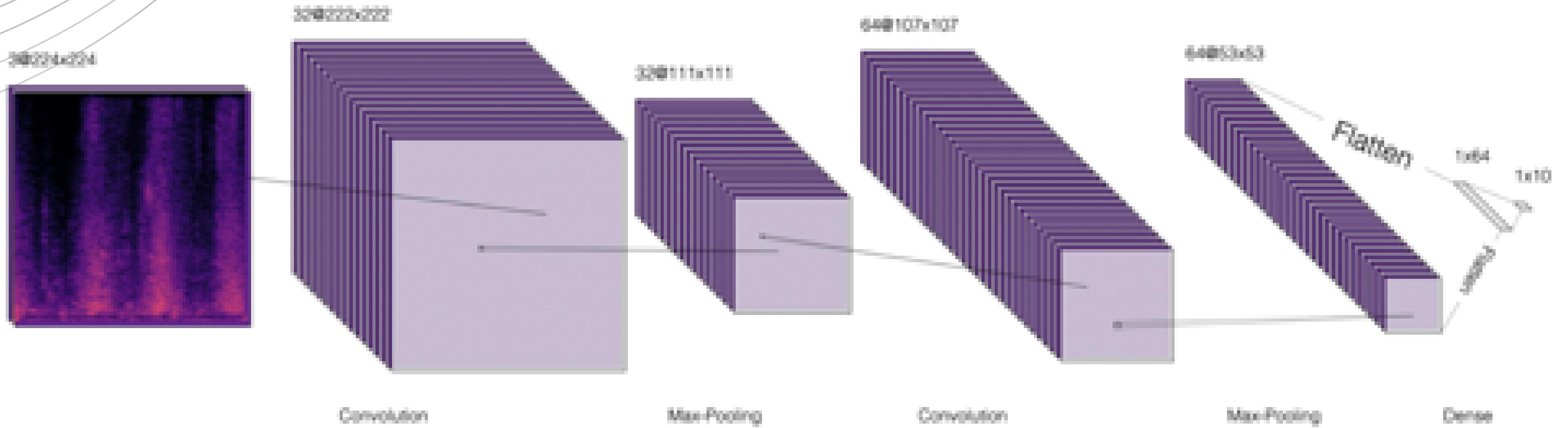
- L1 regularizer (0.001)
- Dropout (0.2), (0.3)
- 39 epochs

# PERFORMANCE



Accuracy	Size of model (KB)	Inference time (s)
94.81	1100.95	0.28

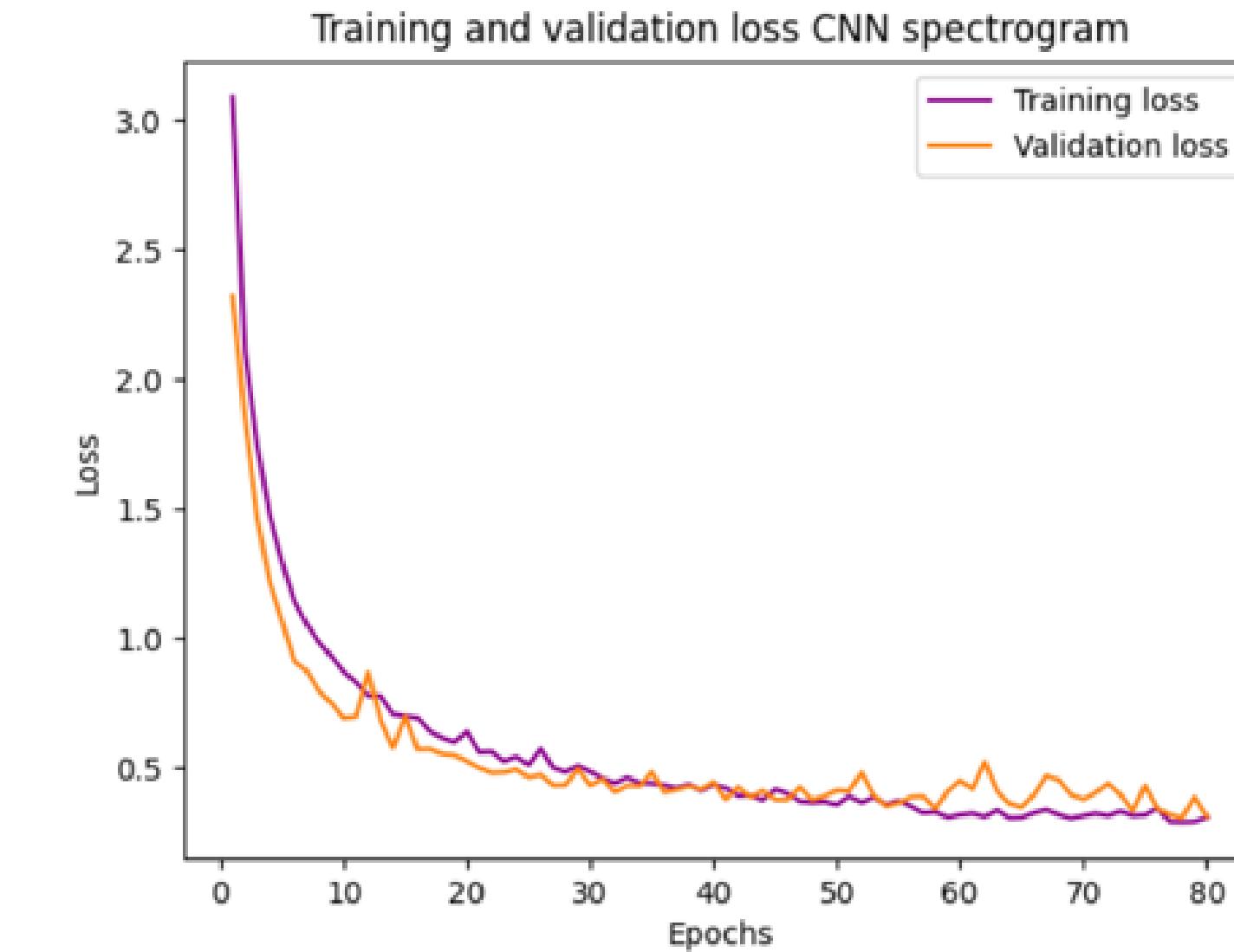
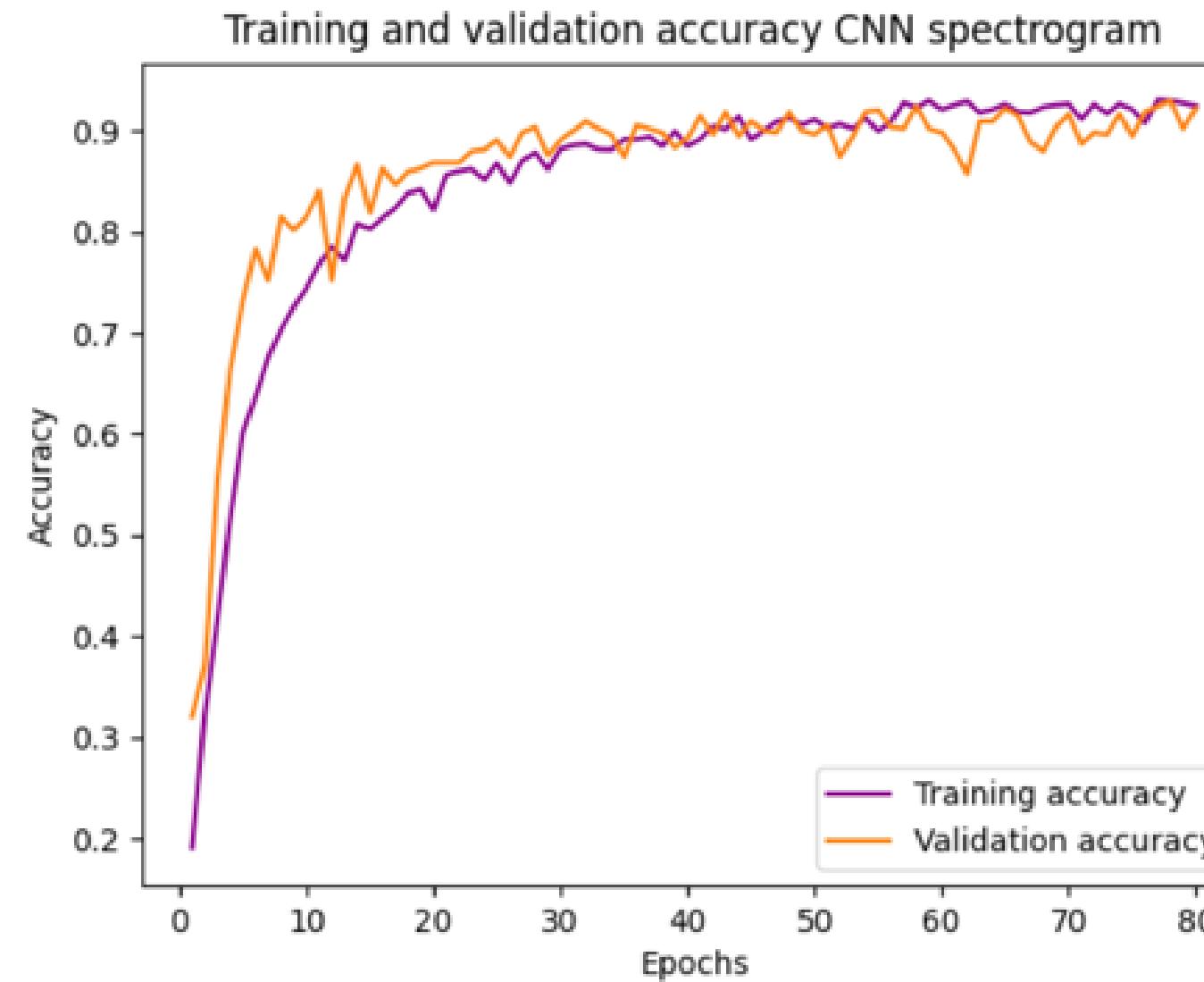
# CNN ON SPECTROGRAM



Moreover, the use of:

- L1 regularizer (0.01)
- Dropout (0.4)
- 80 epochs

# PERFORMANCE

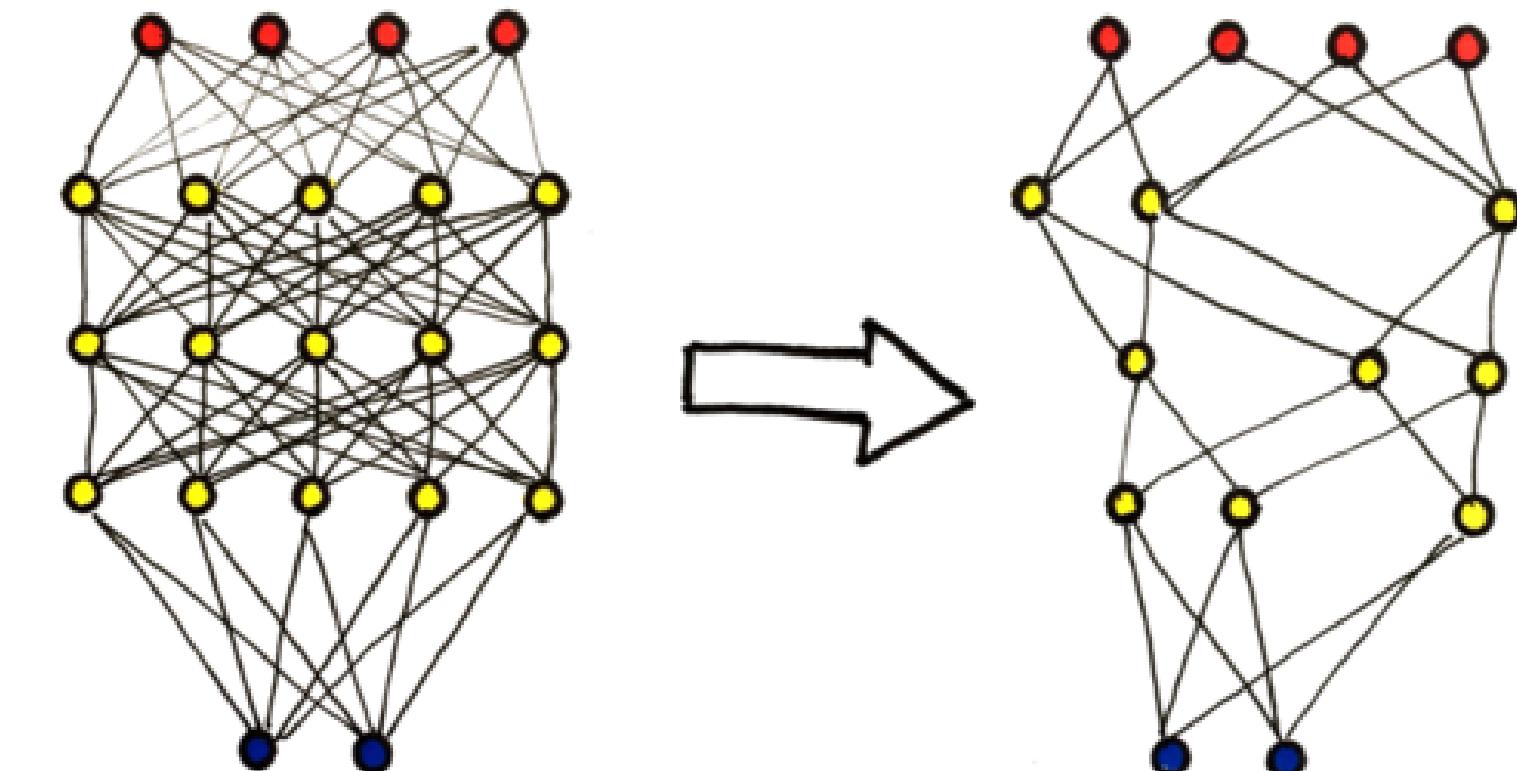


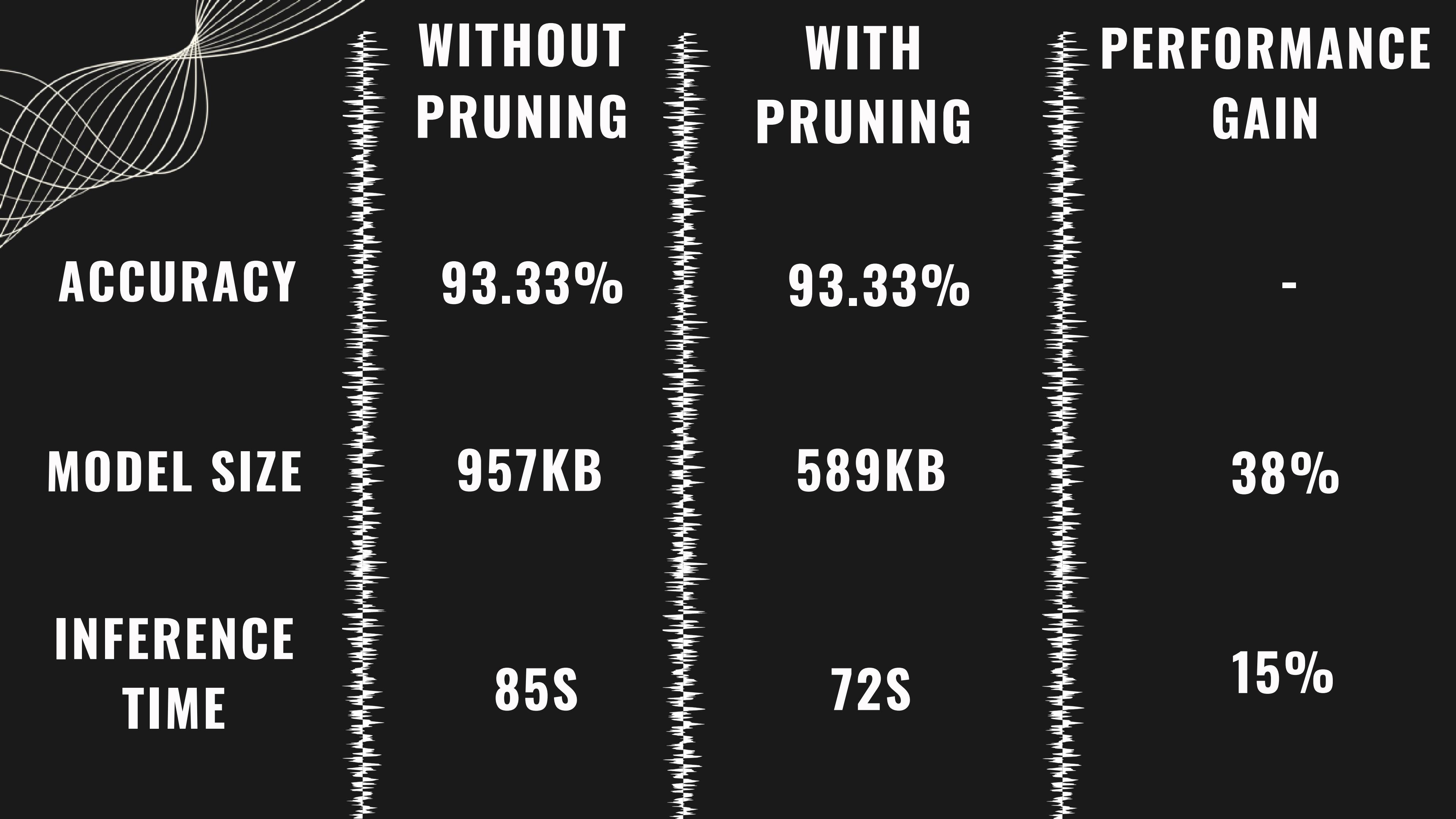
Accuracy	Size of model (KB)	Inference time (s)
93.33	957.85	0.85

# PRUNING

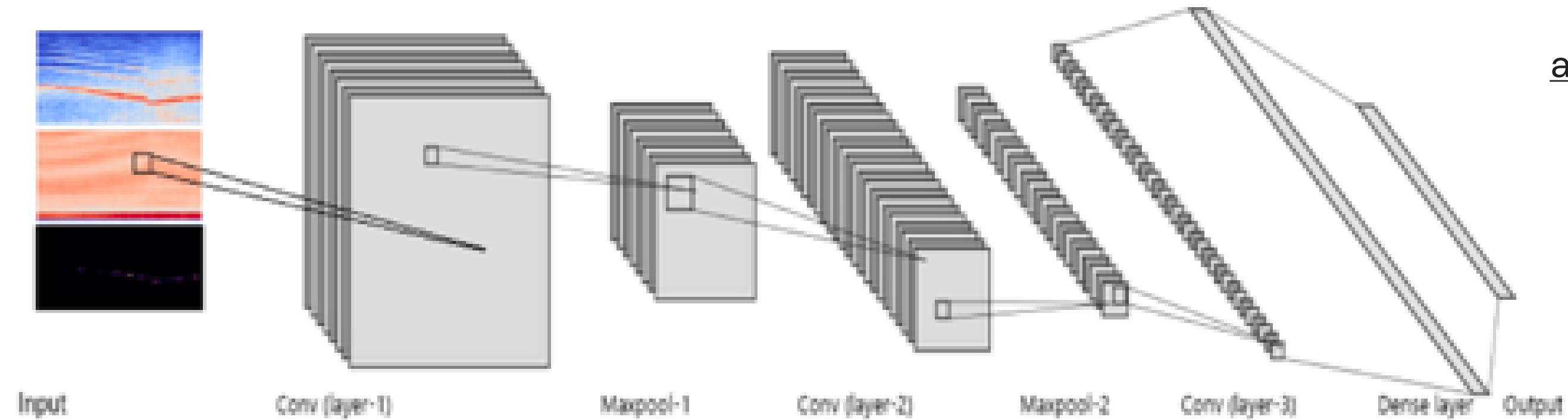
Optimization technique to remove unnecessary connections, parameters, or filters from the network.

- reducing model size
- computational complexity reduction
- maintaining performance.





# CNN ON SPECTROGRAM: PAPER COMPARISON

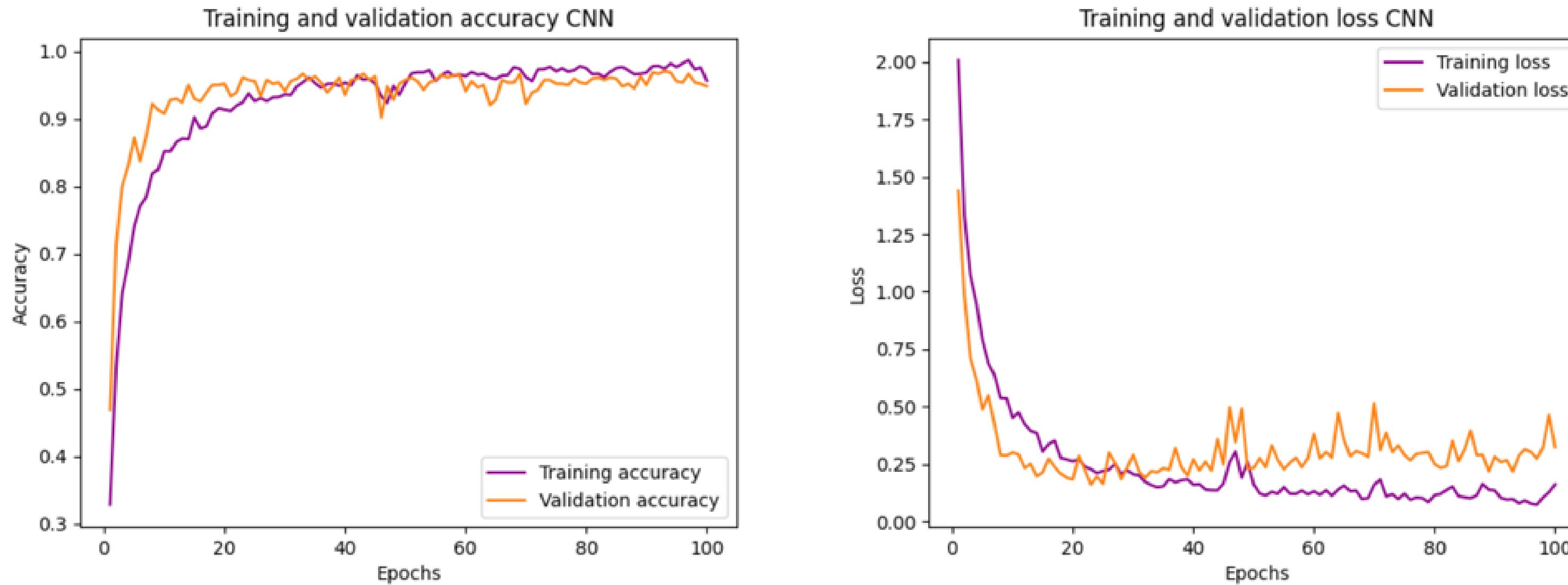


REFERENCE:

Z. Mushtaq and S. F. Su,  
"Environmental sound classification using a regularized deep convolutional neural network with data augmentation", 2020.

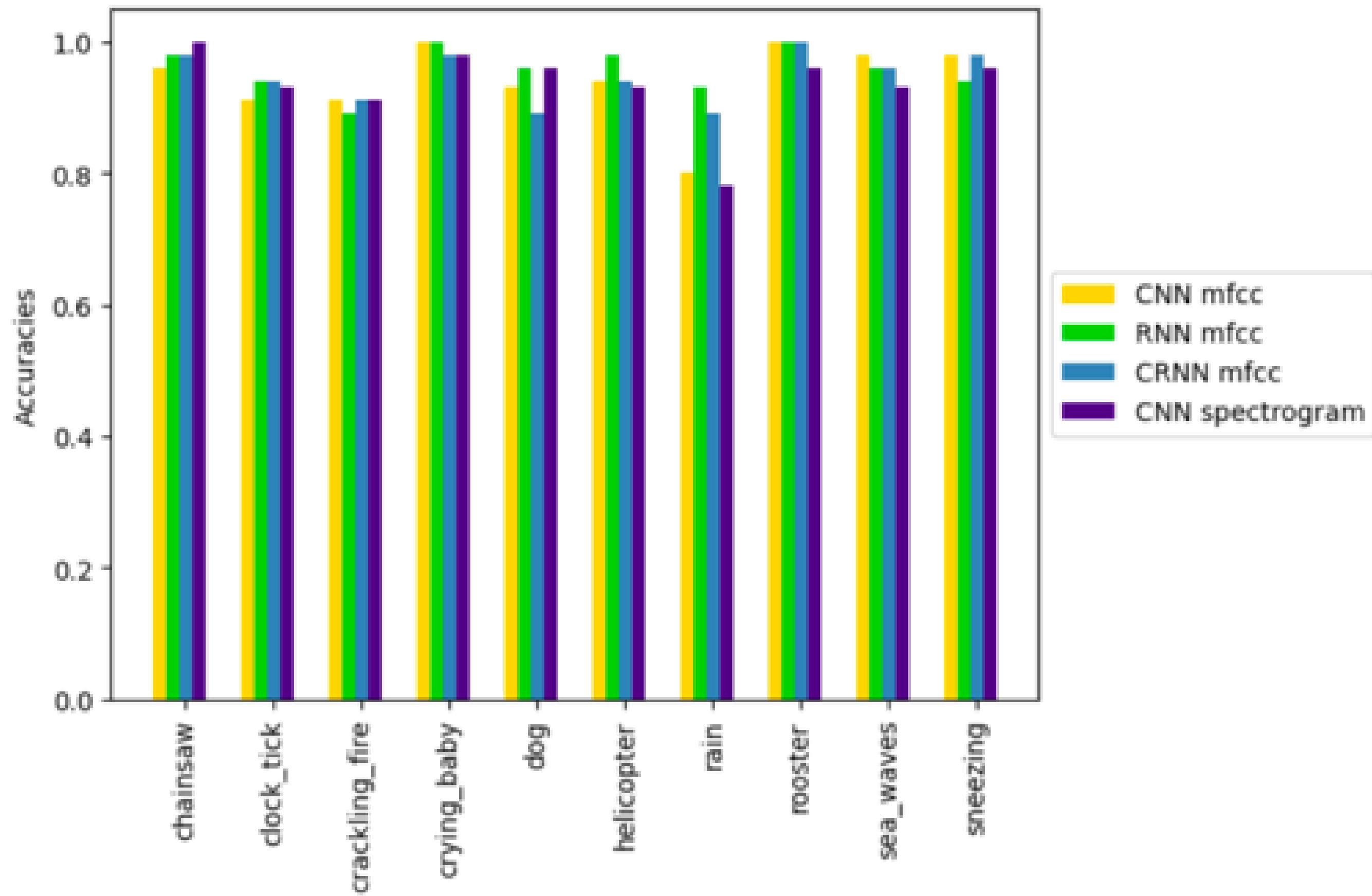
Similarities	Differences
<b>CNN architecture</b> <b>Use of Drop-out</b> <b>Use of Max-Pooling</b> <b>Activation functions</b> <b>Adam and cat-crossentropy</b>	<b>1 less Conv Layer</b> <b>greater Drop-out (0.5)</b> <b>different pool-sizes</b> <b>larger filter sizes</b> <b>fewer tot. parameters</b>

# PERFORMANCE



Accuracy	Size of model (KB)	Inference time (s)
92.96	2545.88	0.36

# FINAL RESULTS.



# FINAL RESULTS.

Input data	Model	Accuracy	Size of model (KB)	Inference time (s)
MFFC	CNN	94.07	1622.69	0.28
MFFC	RNN	95.93	1334.07	0.2
MFFC	CRNN	94.81	1100.95	0.28
Spec	CNN w/o prun	93.33	957.85	0.85
Spec	CNN w/ prun	93.33	589.68	0.72
MFCC	CNN(paper)	92.96	2545.88	0.36

# CONCLUSION



**THANK YOU FOR  
YOUR ATTENTION :)**

**CHIARA DE LUCA - 2039823  
FLAVIA GIANFRATE - 2038501**

