Politecnico di Torino
Machine Learning For IoT

a.a. 2022/2023

Emotion Recognition From Speech

Scalenghe Luca s303452
Feraud Elisa s295573

Table Of Contents

Are we humans good at this?

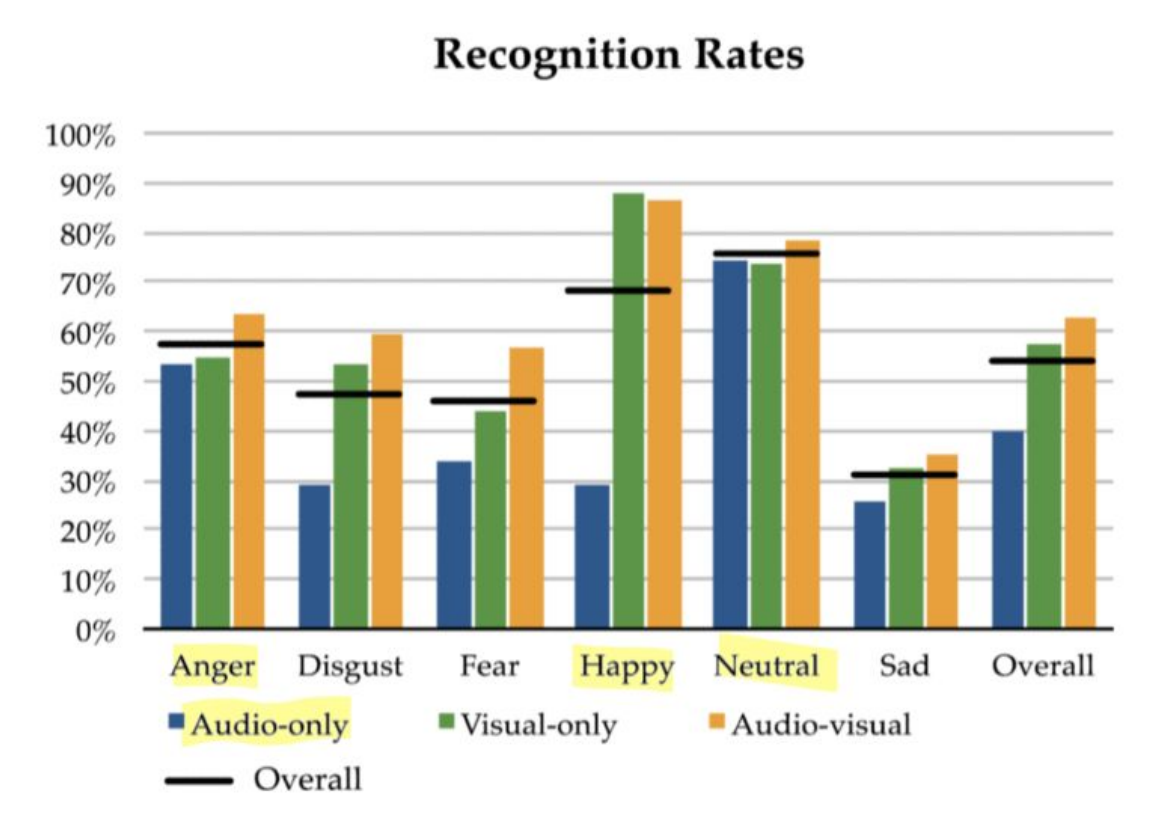
Are we humans good at this?

HAPPY

NEUTRAL

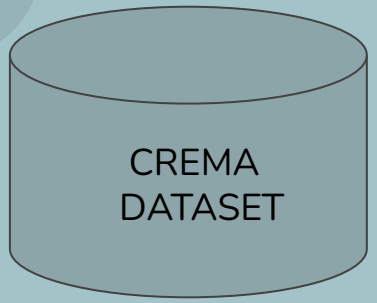
ANGER

How much are people able to recognize emotions?



CREMA-D: Dataset Exploration

CREMA-D: Dataset Exploration



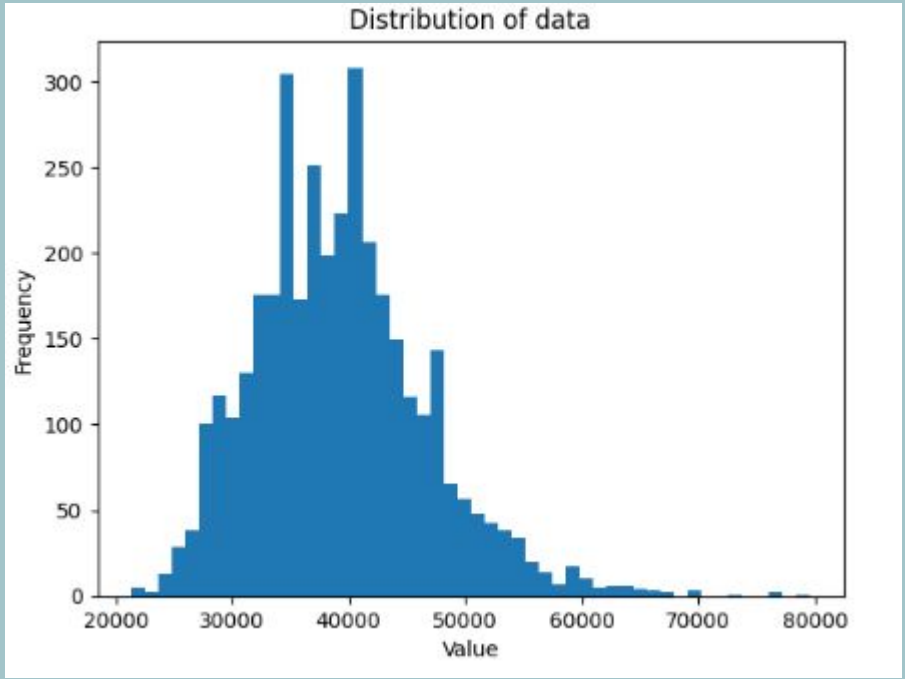
Original Dataset:

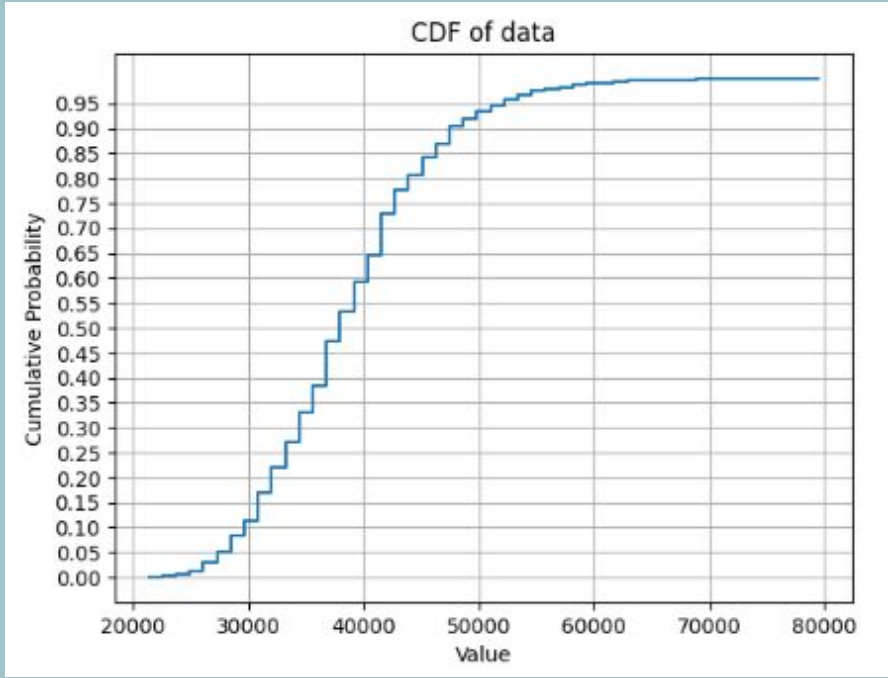
- ➔ 7.442 clips
- ➔ 91 actors
- ➔ 12 sentences
- ➔ 6 emotions: Anger (ANG), Disgust (DIS), Fear (FEA), Happy (HAP), Neutral (NEU), Sad (SAD)
- ➔ 4 emotion levels

Reduced Dataset considered for the project:

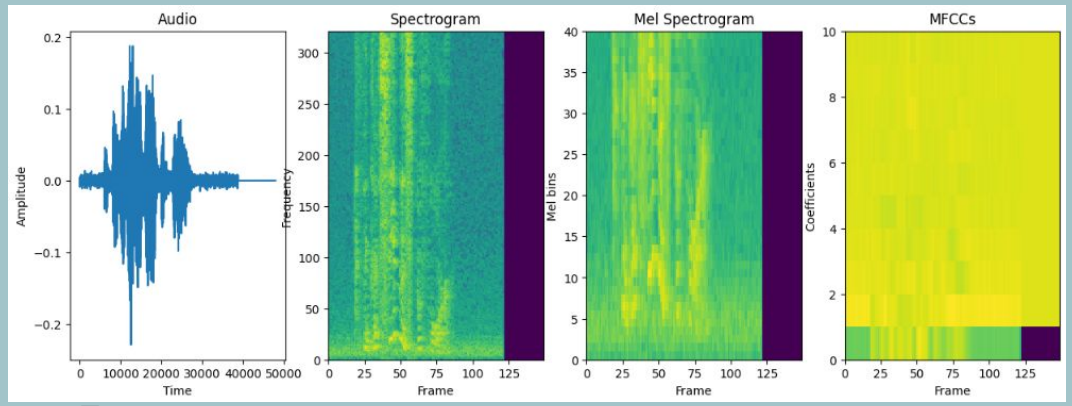
- ➔ 3629 clips
- ➔ 91 actors
- ➔ 3 emotions: Anger (ANG), Happy (HAP), Neutral (NEU)

CREMA-D: Dataset Exploration Lenghts



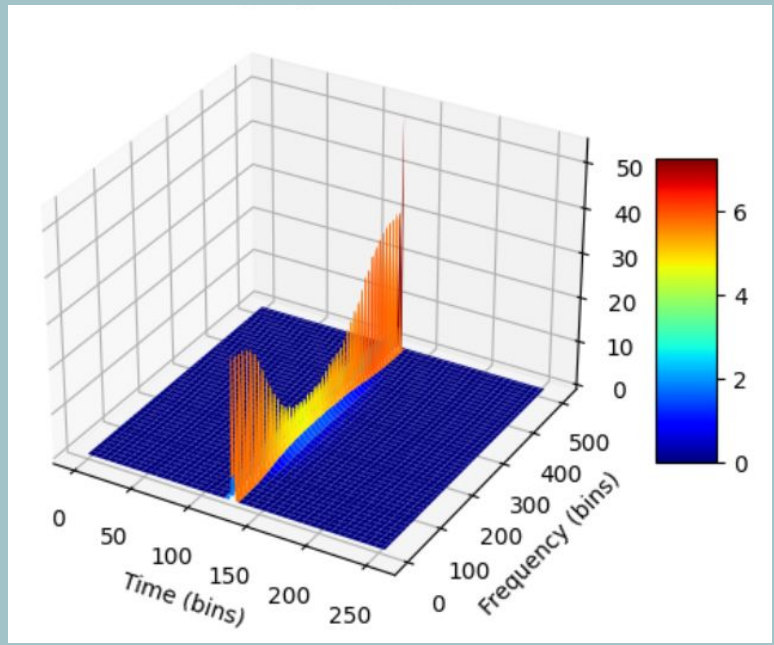


CREMA-D: Dataset Exploration-Happy

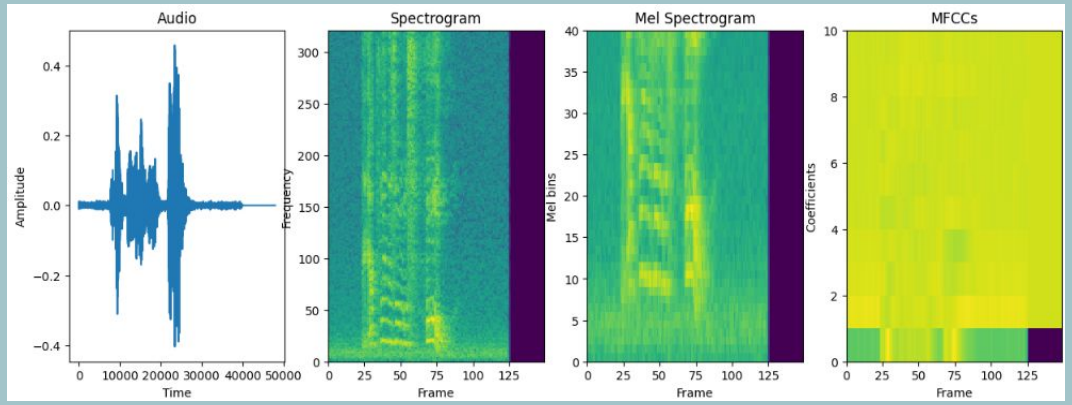


Visualization of a single audio

Average Amplitude - 3D Representation

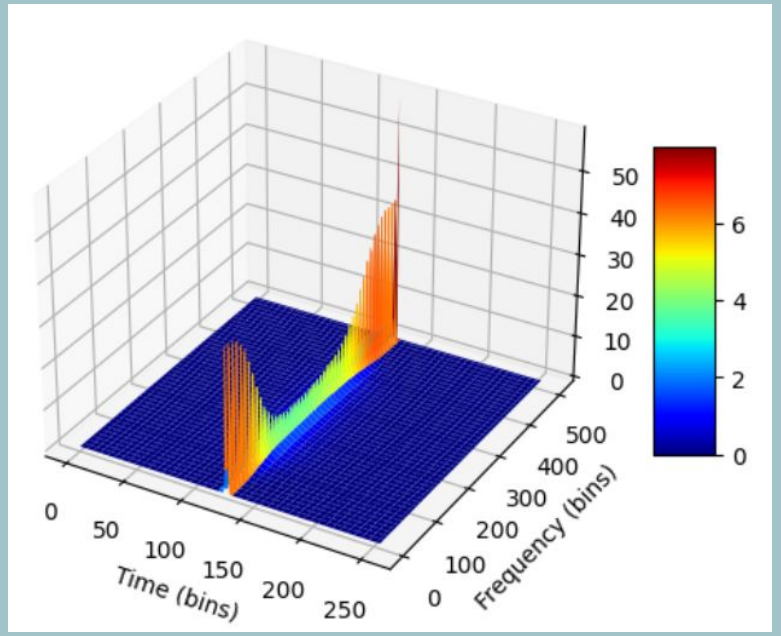


CREMA-D: Dataset Exploration-Anger

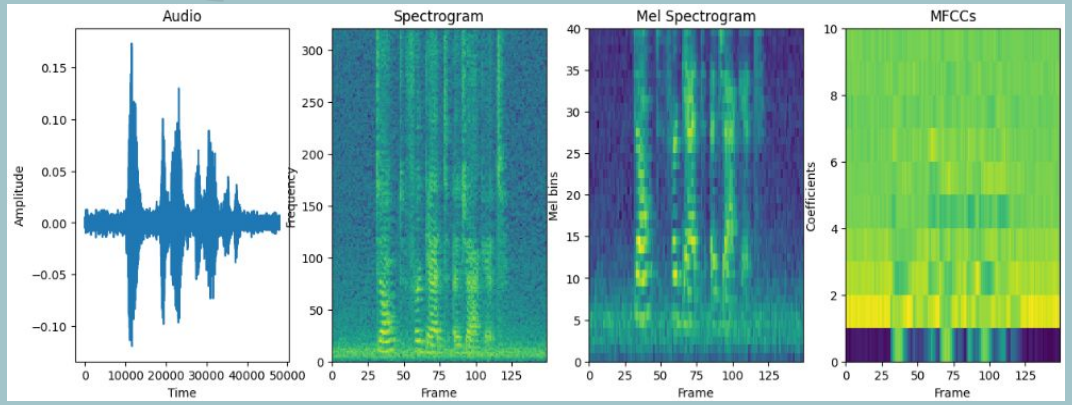


Visualization of a single audio

Average Amplitude - 3D Representation

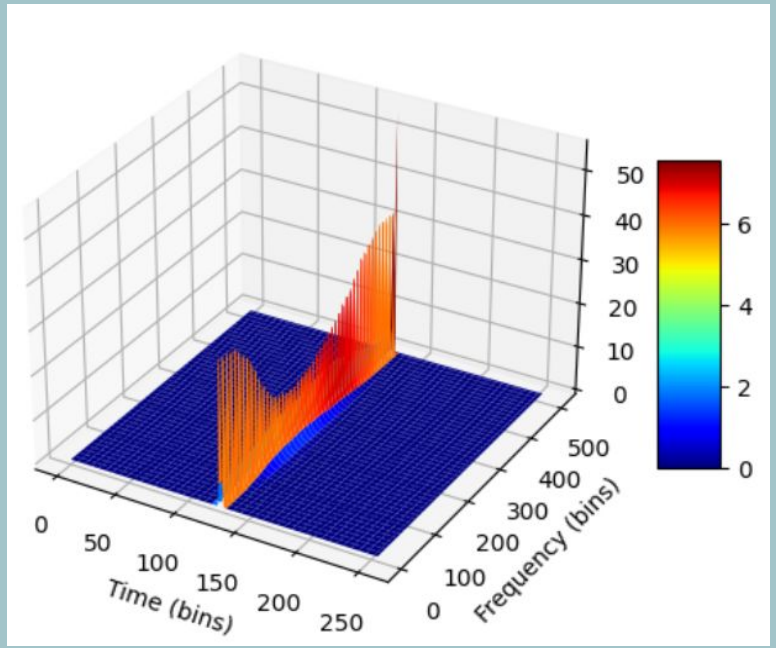


CREMA-D: Dataset Exploration-Neutral



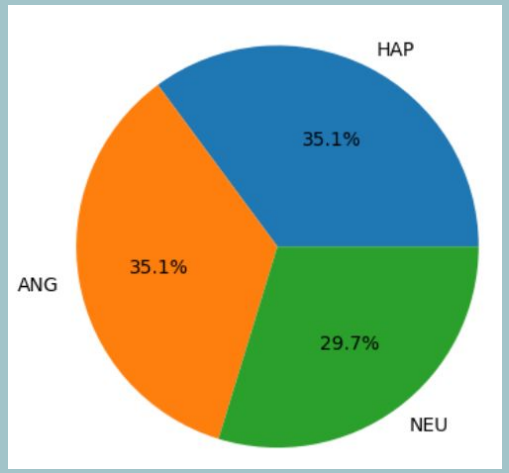
Visualization of a single audio

Average Amplitude - 3D Representation

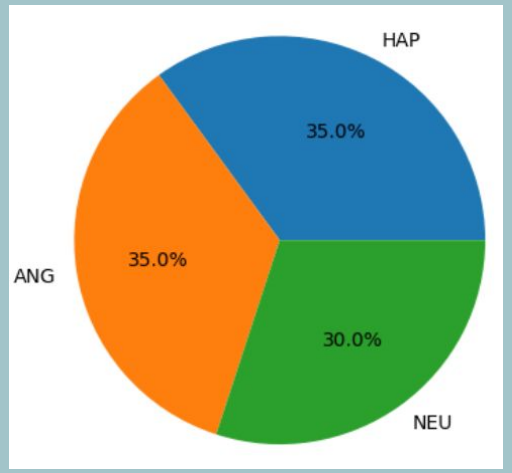


CREMA-D: Dataset Exploration

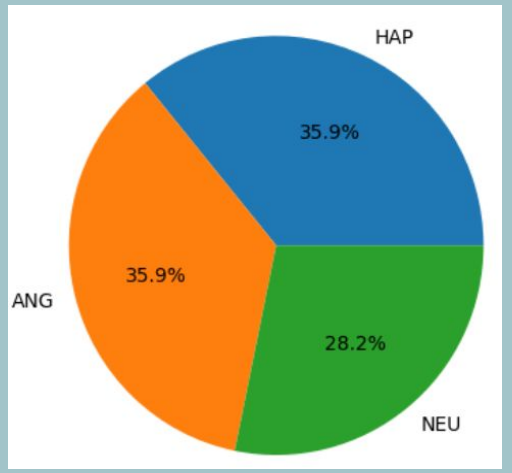
Number of happy, anger and neutral audios for each actor.



Example 1



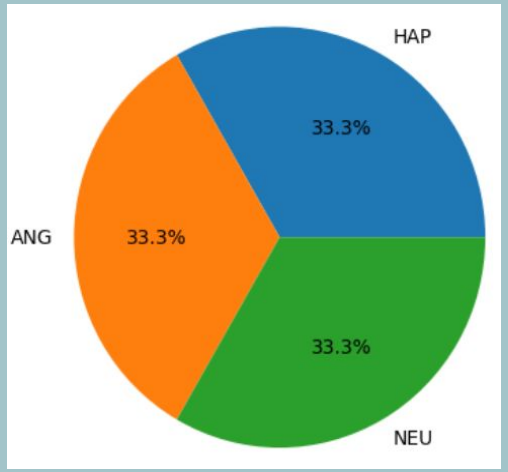
Example 2



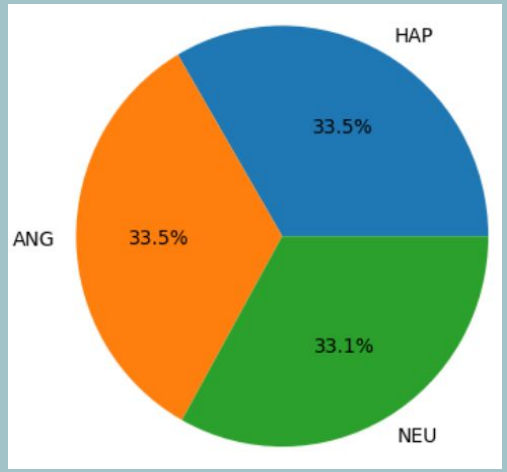
Example 3

CREMA-D: Dataset Exploration

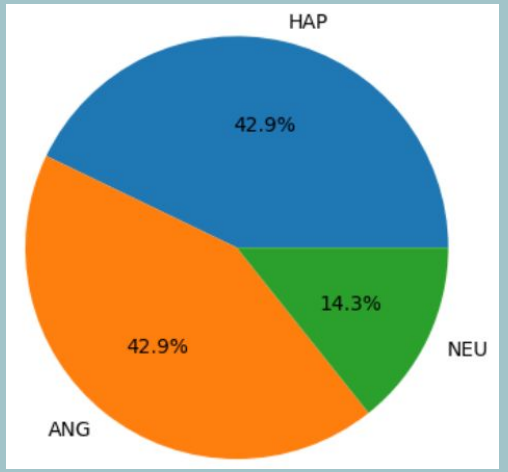
Number of happy, anger and neutral audios for each sentence.



Example 1



Example 2



Example 3

TF-LITE MODEL

The model

- 5 Convolutional layers
- Combination of three optimization techniques:
  - The Width pruning to reduce the filter size.
  - The Weight pruning in order to prune weights
  - The Depth-wise separable convolution in order to reduce the number of parameters in the model and make it more efficient to compute.

The model - Metrics

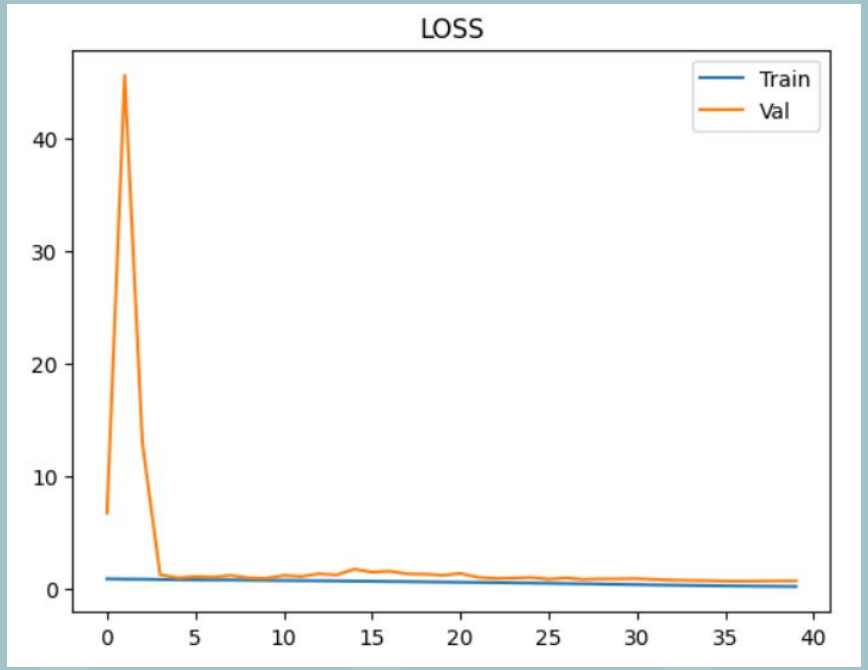
Loss:

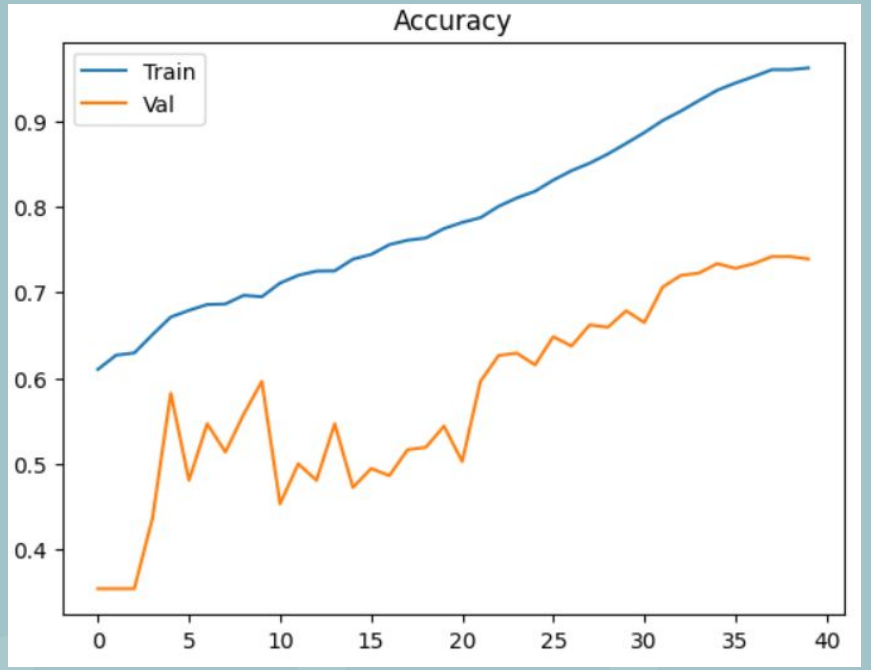
- Sparse Categorical Cross Entropy
- It increases as the predicted probability diverges from the actual label

Accuracy:

- Sparse Categorical Accuracy
- It calculates how often predictions match integer labels

Training and Validation sets





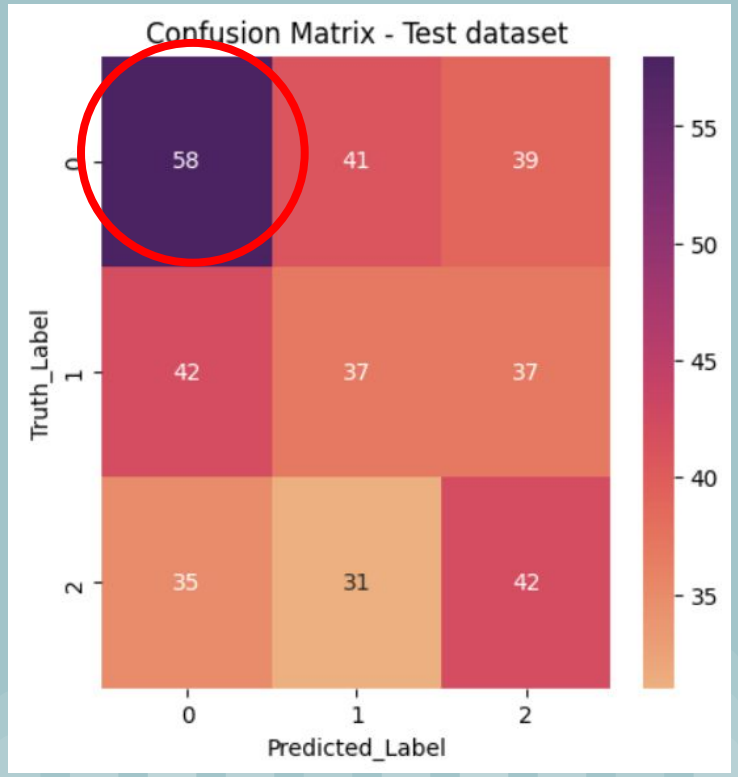
RESULTS

How much is our model able to recognize emotions?

ANG

HAP

NEU

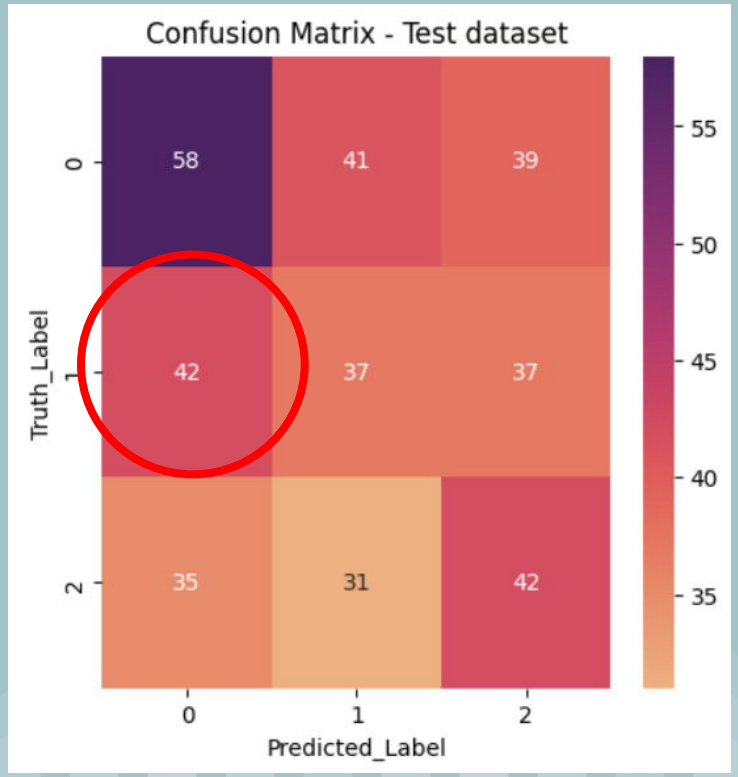


How much is our model able to recognize emotions?

ANG

HAP

NEU

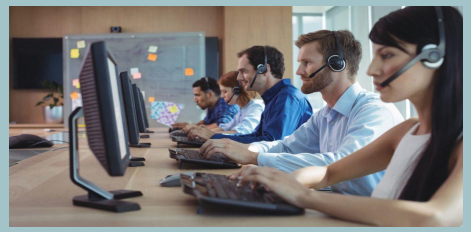


Different models comparison

Test accuracy	Zipped TFlite size	Latency median	Layers	Alpha	Final sparsity	Dataset
79,6%	288.4KB	15.0ms	5 (all 256)	0,5	0,5	Balanced
78,2%	197.5KB	8.6ms	5 (256,256, 128,192,256)	0,5	0,5	Balanced
74,9%	197.5KB	8.6ms	5 (256,256, 128,192,256)	0,5	0,5	Augmented
73,8%	120.079 KB	9.5ms	5 (256,256, 128,192,256)	0,5	0,5	Balanced
71,5%	105.234 KB	8.7ms	5 (256,256, 128,192,256)	0,5	0,5	Augmented

Possible use case scenarios

Education



Call centers

Personal assistants

Advertising

COMMUNICATION

Message broker & Server

Message broker

- Decoupling of the communication
- Isolate potential problems
- Prevents from security vulnerabilities

Server

- Devices(): returns all the monitored devices
- SingleDevice(): returns informations of a Single Device
    - mac_address
    - labels
    - timeseries_ANG
    - timeseries_NEU
    - timeseries_HAP
- Delete(): deletes all the data stored on the server

Client

HTTP protocol for communication with server

Statistics plotted:

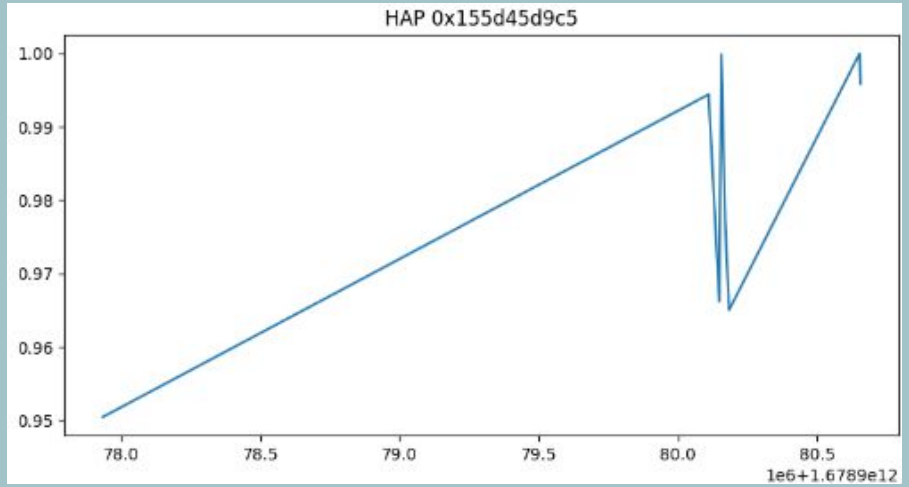
- Line plot: confidence value over time for each emotion
- Histogram: frequency of confidence interval for each emotion
- Mean confidence in each emotion and the mean confidence over all:

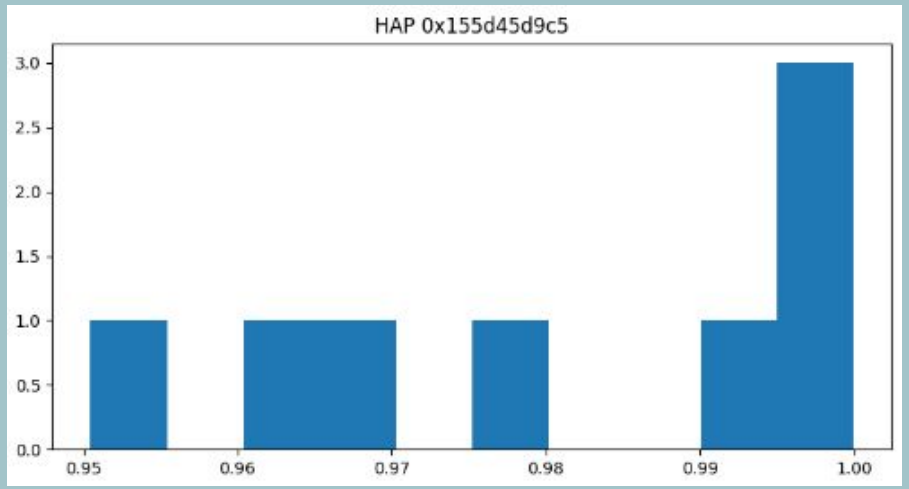
The mean confidence in ANG is 0.992

The mean confidence in NEU is 0.991

The mean confidence in HAP is 0.981

The mean confidence overall is 0.990





Thanks for the attention!