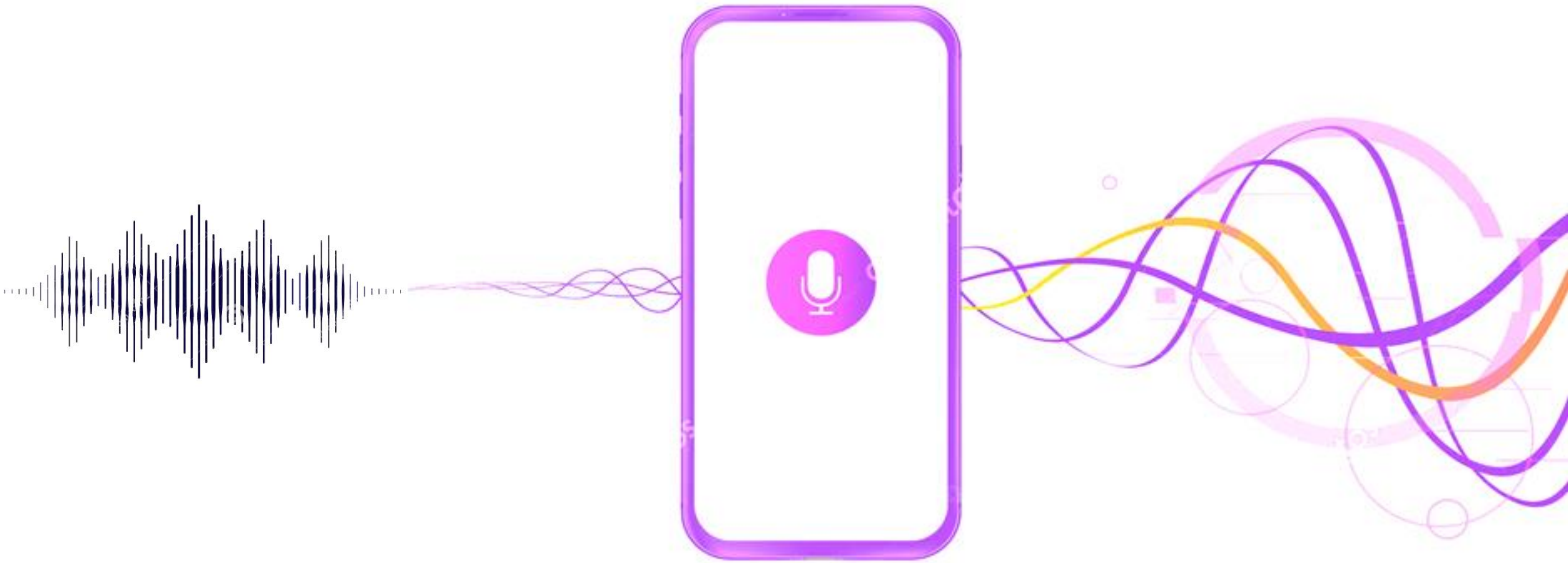
PUC RIO

Project

# A Real-time ENVIRONMENTAL SOUND RECOGNITION SYSTEM using Machine Learning Techniques.

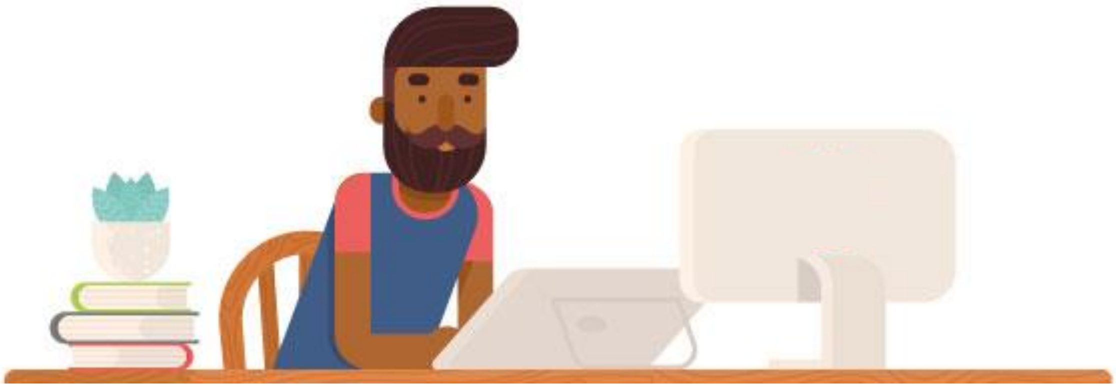
Asiel Aldana Ortiz

# Motivation:



✓ SOUND RECOGNITION for environmental sounds classification

# Objective:



- Obtain a model based on a **CONVOLUTIONAL NEURAL NETWORK** for environmental sound classification.

- Develop a **DISTRIBUTED(or EMBEDDED) COMPUTING ARCHITECTURE** for real-time classification of environmental sounds.

# URBAN SOUND DATASET

| class | Samples | Duration (avg) | In foreground |
|---|---|---|---|
| air_conditioner | 1000 | 3.99 s | 56 % |
| car_horn | 429 | 2.46 s | 35 % |
| children_playing | 1000 | 3.96 s | 58 % |
| dog_bark | 1000 | 3.15 s | 64 % |
| drilling | 1000 | 3.55 s | 90 % |
| engine_idling | 1000 | 3.94 s | 91 % |
| gun_shot | 374 | 1.65 s | 81 % |
| jackhammer | 1000 | 3.61 s | 73 % |
| siren | 929 | 3.91 s | 28 % |
| street_music | 1000 | 4.00 s | 62 % |

- 8732(Annotations+Records)
- <= 4s
- 10 Class

www.freesound.org

Environment Sound

ESC-50 Dataset

- 2000 (lb+rec)

Freesound.org

Urban Sound Dataset

- 8732(lb+rec)
- <=4s
- 10 class

AudioSet Dataset

- 632 class(lb+rec)

Bird Sound

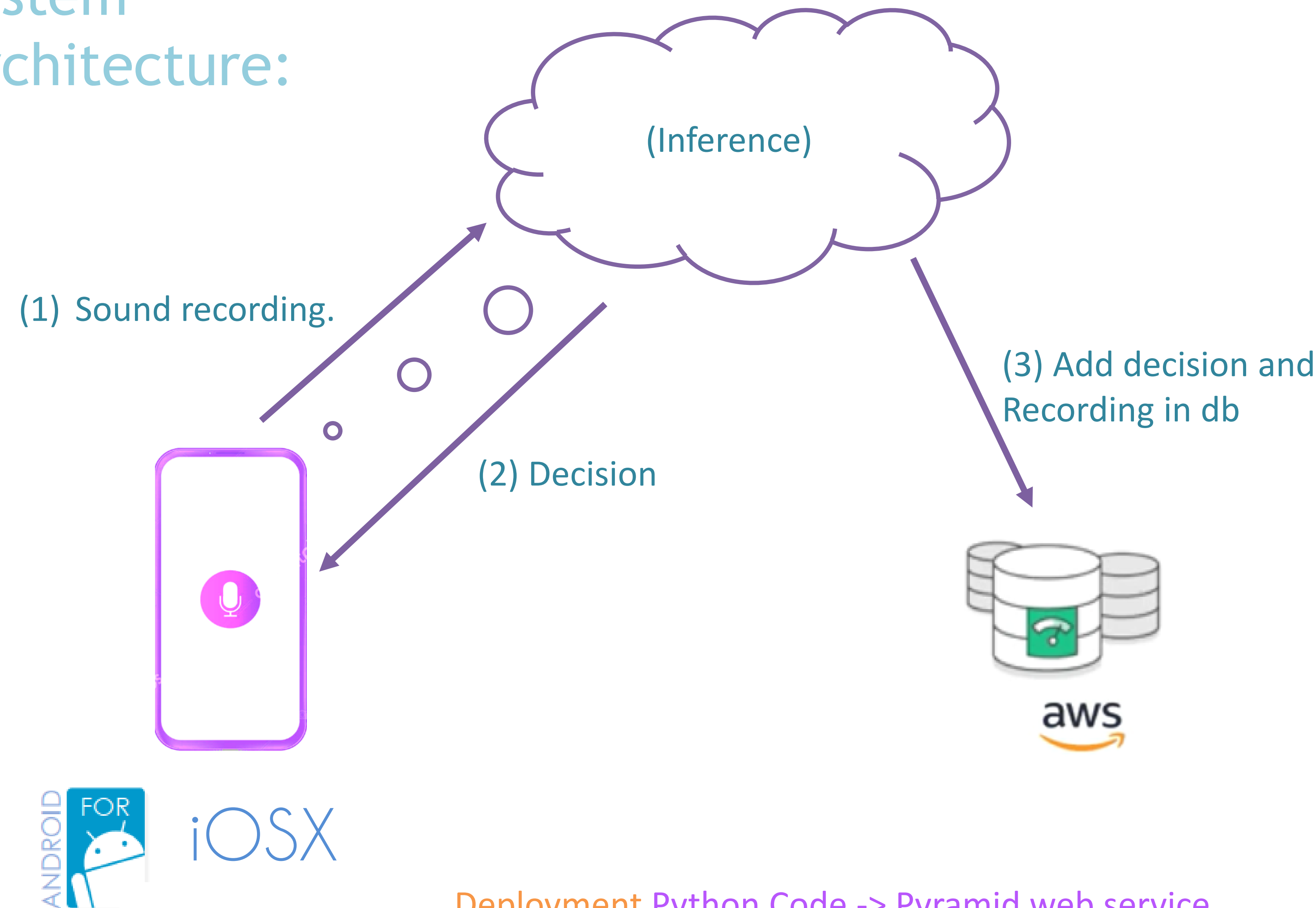
warblrb10k

- 8000 (lb+rec)

Bird Sounds

- T-SNE

# System Architecture:




Deployment Python Code -> Pyramid web service

# Working plan:

1. MACHINE LEARNING TECHNIQUES.
   - Classification.
   - Training process.

2. NEURAL NETWORKS
   - Multi-Layer Perceptron.
   - Convolutional Neural Network(CNN).
   - Convolutional Recurrent Neural Network.
   - Efficient CNNs for Image Classification.

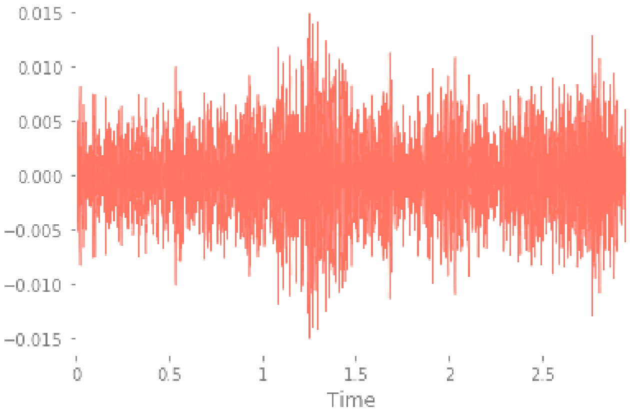
3. AUDIO CLASSIFICATION
   - Digital sound.
   - Audio Signal Preprocessing Techniques.
     - Spectrogram.
     - Mel-spectrogram.
     - Data augmentation.
     - Normalization.
     - Analysis Windows.
   - Weak labeling.
   - Efficient CNNs for Sounds Classification.

4. ENVIRONMENTAL SOUND RECOGNITION(ESR)
   - Datasets.
   - Feature Extraction.
     - Stationary “ESR” Techniques:
       - Zero-crossing Rate (ZCR).
       - Short-time Energy (STE)
       - MFCCs
     - Non-stationary “ESR” Techniques:
       - Wavelet-based methods.
       - Power-spectrum-based methods.
       - Sparse-representation-based methods.

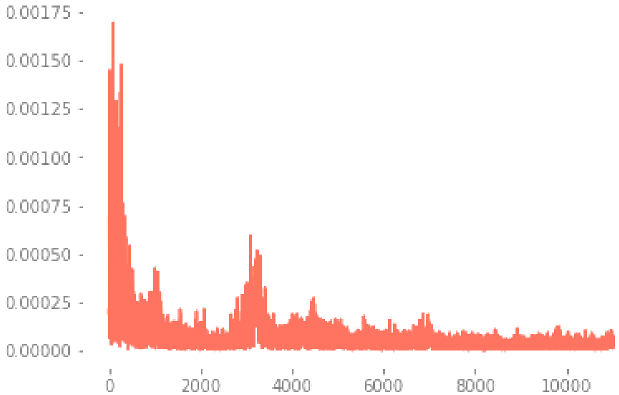
5. DATABASE
   - Warblrb10k.
   - Bird Sounds
   - Urban Sound Dataset.

6. COMPUTATIONAL ARCHITECTURE
   - Multilayer Computing Architecture.
   - Audio Streaming.

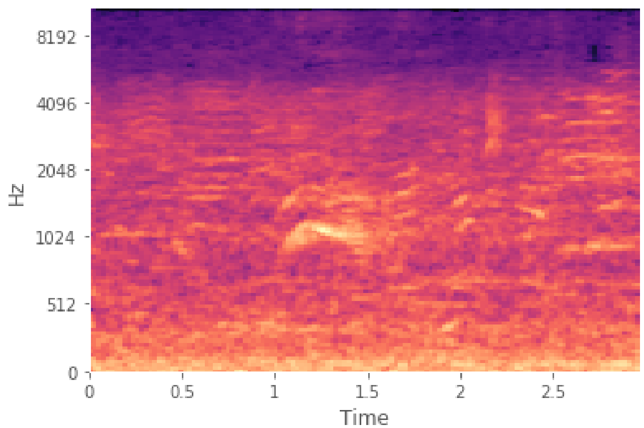
# SIGNAL REPRESENTATION
(Feature extraction)




♪ Amplitude/Time



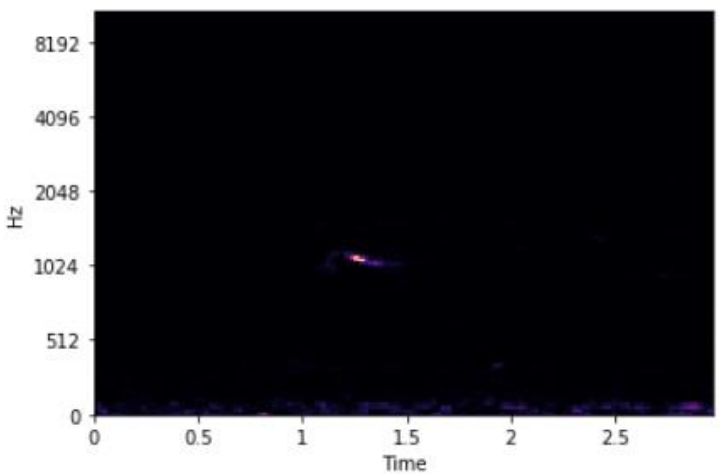

♪ Frequency/Time




♪ Amplitude/Frequency

# SIGNAL REPRESENTATION
# (Feature extraction)

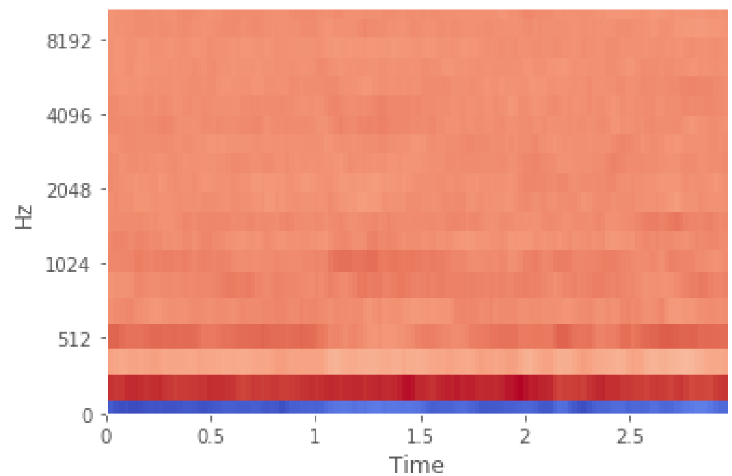
100263-2-0-121.wav




**Mel Frequency Cepstrum(MFC):**

Is a representation of
the Short-Term Power
Spectrum of a sound.

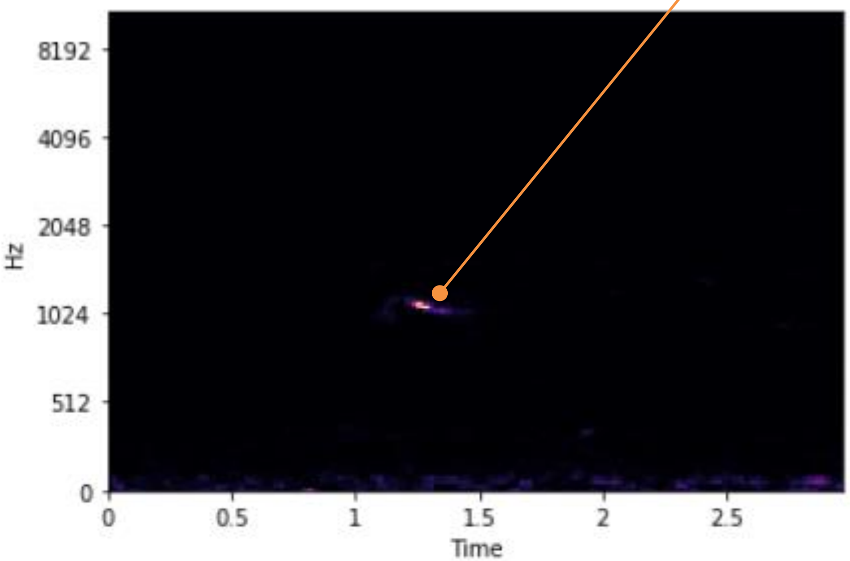
100263-2-0-121.wav(20 Coefficients per frame)



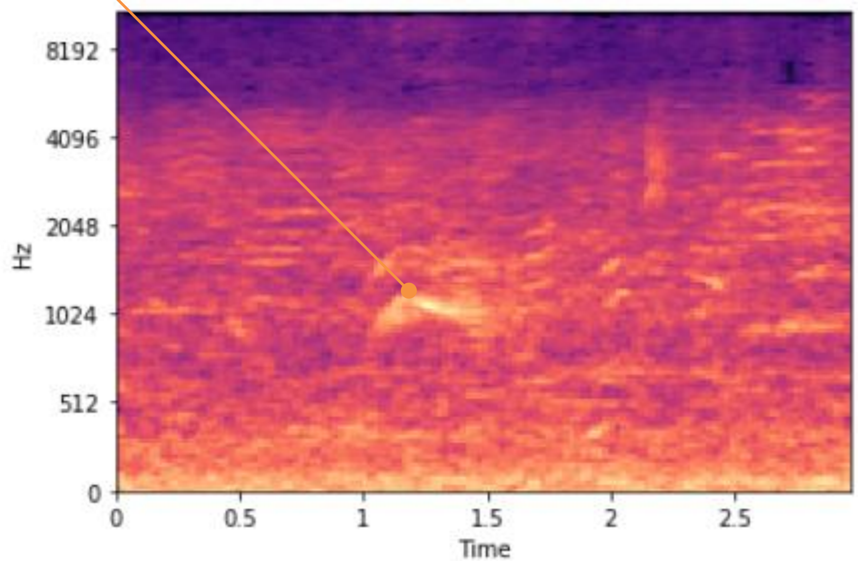

**Mel Frequency Cepstral Coefficients(MFCCs)**

Are coefficients that
collectively make up an MFC.

# Power to Decibel

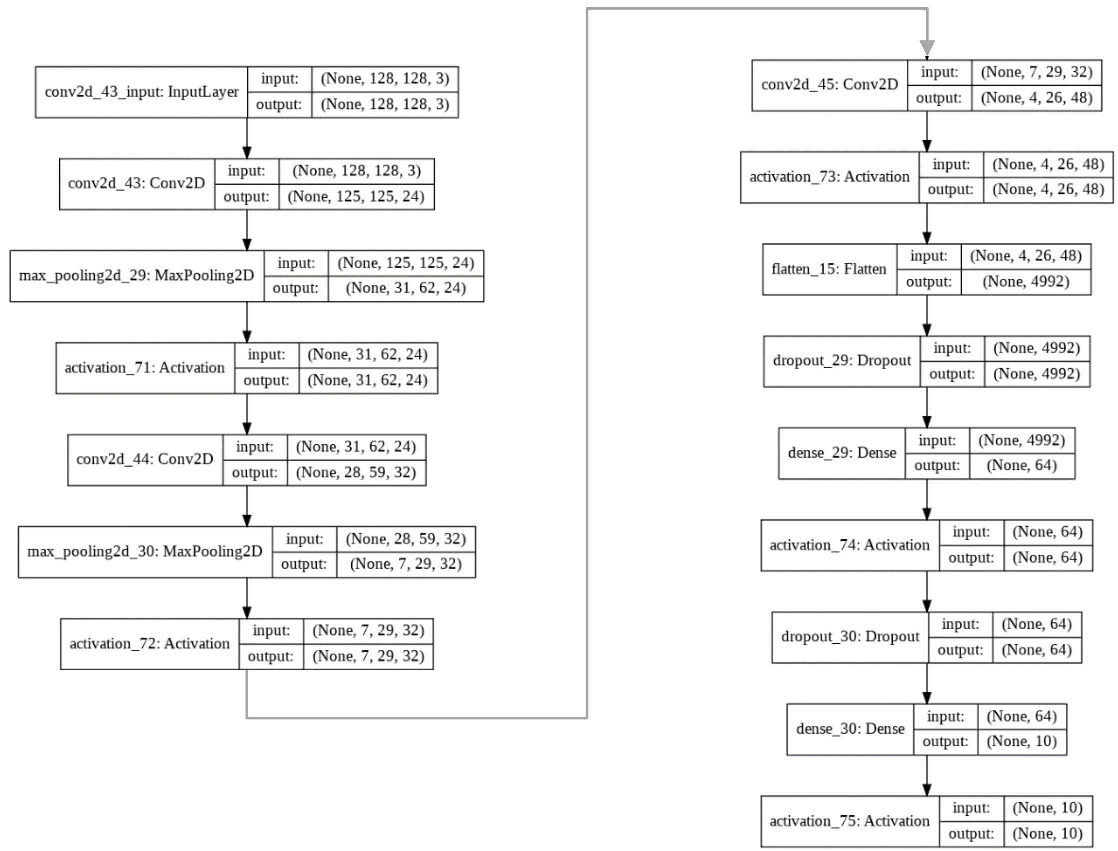


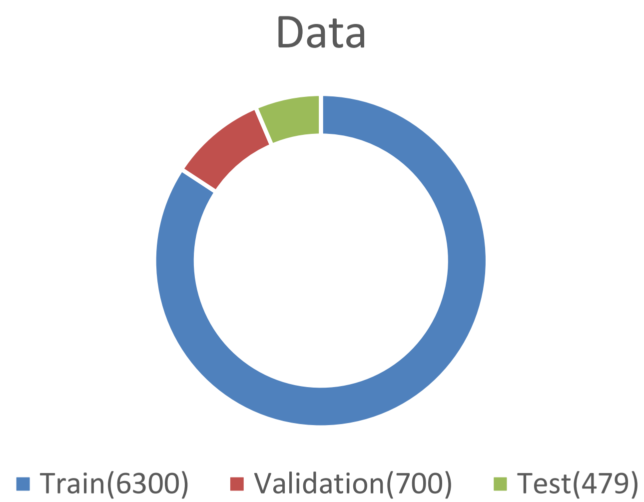

MFC(Power)

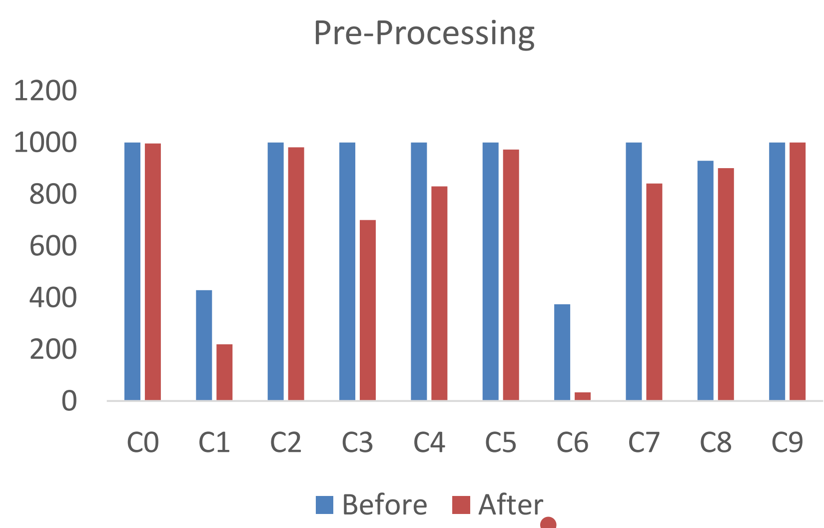



MFC(Decibel)

# MODEL


conv2d_43_input: InputLayer
input: (None, 128, 128, 3)
output: (None, 128, 128, 3)
conv2d_43: Conv2D
input: (None, 128, 128, 3)
output: (None, 125, 125, 24)
max_pooling2d_29: MaxPooling2D
input: (None, 125, 125, 24)
output: (None, 31, 62, 24)
activation_71: Activation
input: (None, 31, 62, 24)
output: (None, 31, 62, 24)
conv2d_44: Conv2D
input: (None, 31, 62, 24)
output: (None, 28, 59, 32)
max_pooling2d_30: MaxPooling2D
input: (None, 28, 59, 32)
output: (None, 7, 29, 32)
activation_72: Activation
input: (None, 7, 29, 32)
output: (None, 7, 29, 32)
conv2d_45: Conv2D
input: (None, 7, 29, 32)
output: (None, 4, 26, 48)
activation_73: Activation
input: (None, 4, 26, 48)
output: (None, 4, 26, 48)
flatten_15: Flatten
input: (None, 4, 26, 48)
output: (None, 4992)
dropout_29: Dropout
input: (None, 4992)
output: (None, 4992)
dense_29: Dense
input: (None, 4992)
output: (None, 64)
activation_74: Activation
input: (None, 64)
output: (None, 64)
dropout_30: Dropout
input: (None, 64)
output: (None, 64)
dense_30: Dense
input: (None, 64)
output: (None, 10)
activation_75: Activation
input: (None, 10)
output: (None, 10)

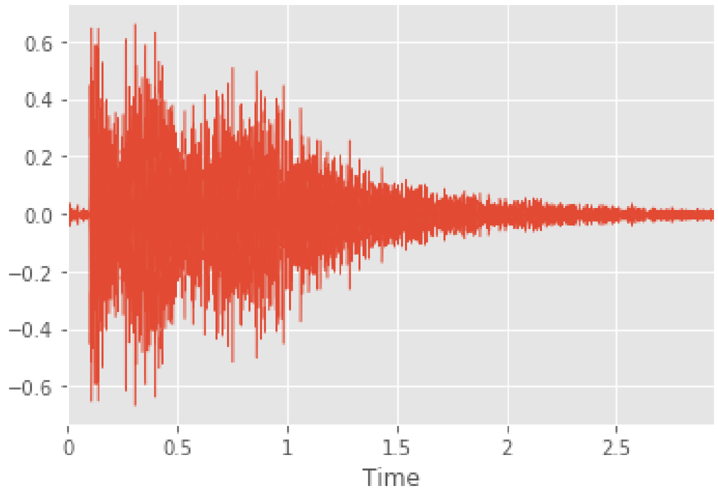


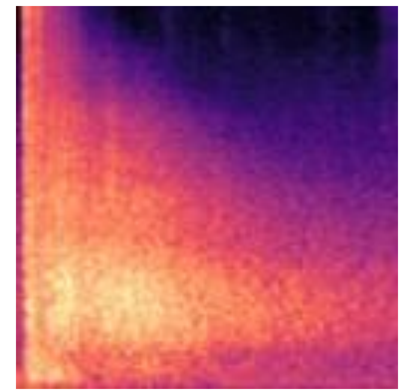

```
Duration_Sound = 2.95 s
sr = 22050 Hz
t(s)= hop_lennght/sr ~23ms
N = 2.95/t
mf = n_mels
shape(N,mf) = 128x128
```

# Data Augmentation

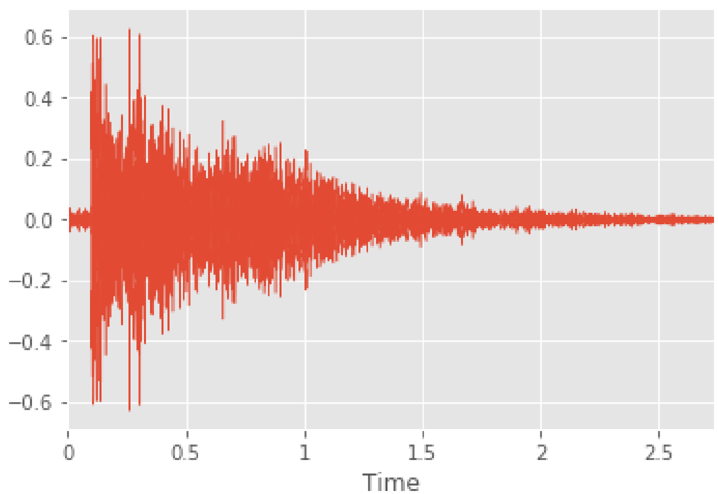
rate = 1.08



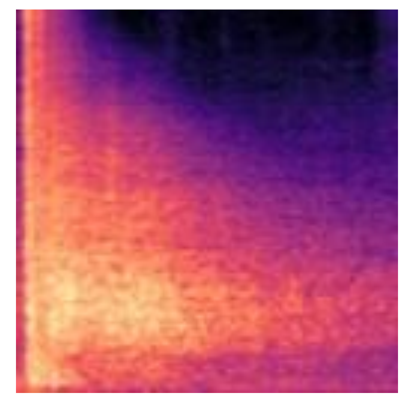

Original



122690-6-0-0.jpg




Obtained



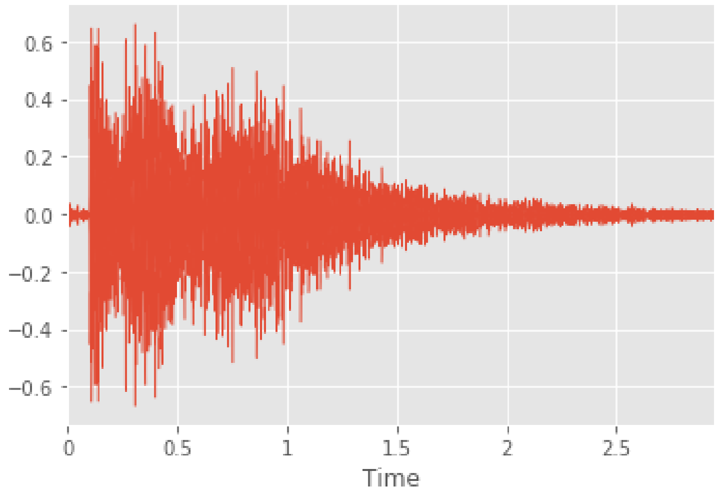
122690-6-0-0TS108_c6.jpg

Time Stretch Class 6 (122690-6-0-0TS108_c6.wav)

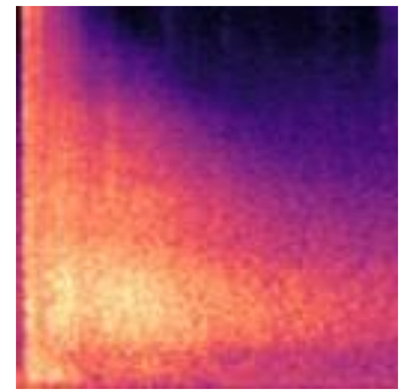
0.82, 0.94 -> 90 increased samples(30/34)

# Data Augmentation

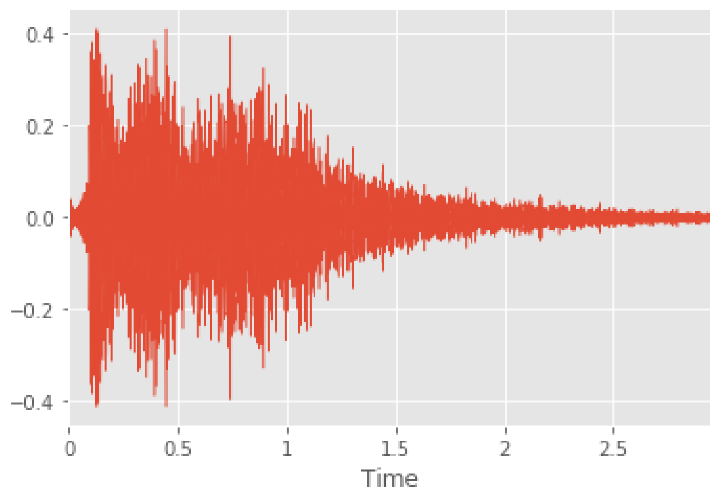
`n_steps = 2`



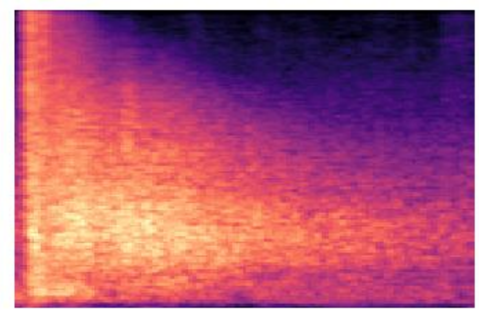

Original



122690-6-0-0.jpg



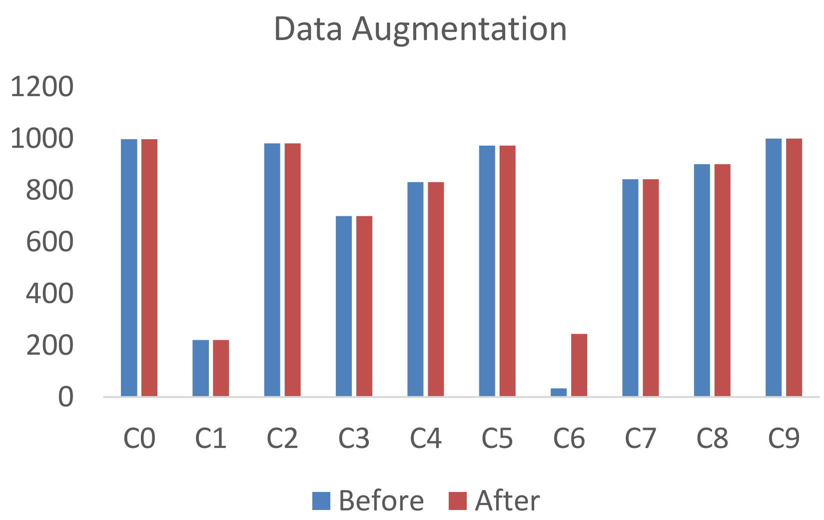

Obtained



122690-6-0-0mfccPS2_c6.wav

Pitch Shift Class 6 (122690-6-0-0mfccPS2_c6.wav)

−2, −1, 1 −>120 increased samples(30/34)



Data Augmentation
1200
1000
800
600
400
200
0
C0
C1
C2
C3
C4
C5
C6
C7
C8
C9
Before
After

# TESTS

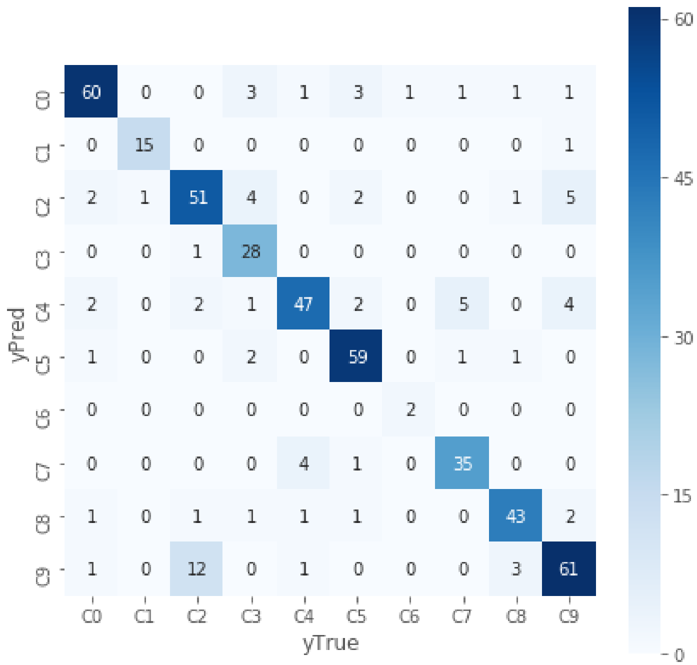
| Learning rate | Epochs | Batch_size | N_Test |
|---|---|---|---|
| 0.01-0.001 | 20-100 | 100-128 | 479 |

| Learning rate | Epochs | Batch_size | N_Test |
|---|---|---|---|
| 0.001 | 55 | 100 | 479 |

```
loss="categorical_crossentropy"
optimizers.Adamax
optimizers.Adam
optimizers.RMSprop
```

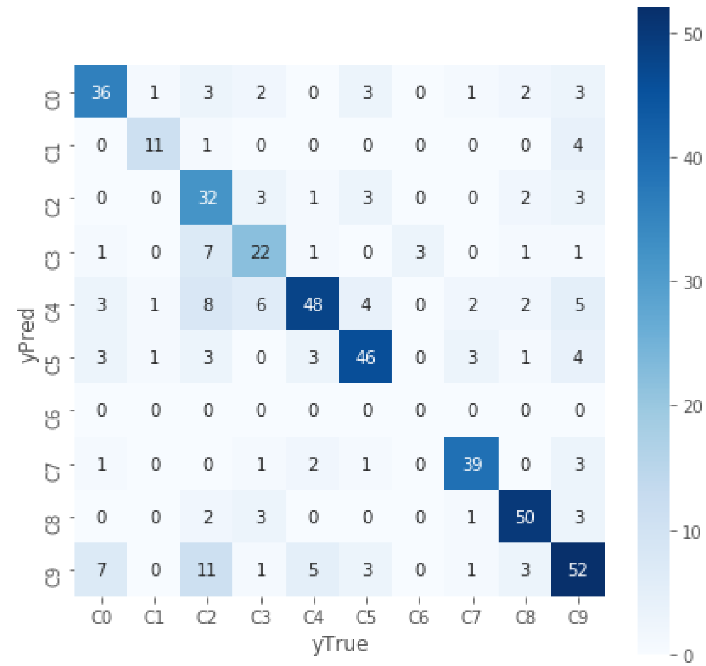
# RESULTS MFC/MFCCs

| Learning rate | Epochs | Batch_size | N_Test |
|---|---|---|---|
| 0.001 | 55 | 100 | 479 |




Confusion matrix MFC

0.80<Acc<= 0.84

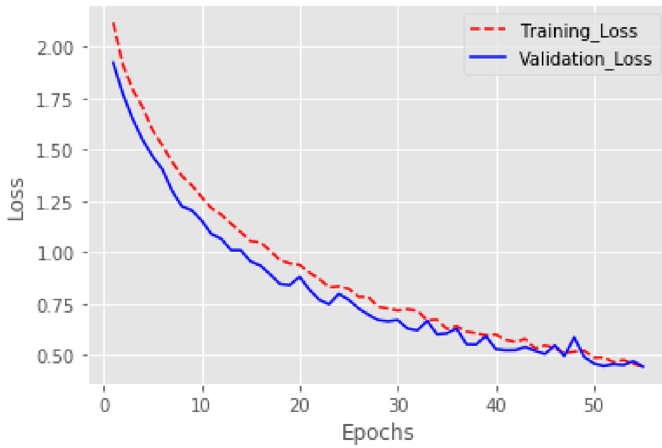



Confusion matrix MFCCs

0.65<Acc<= 0.70

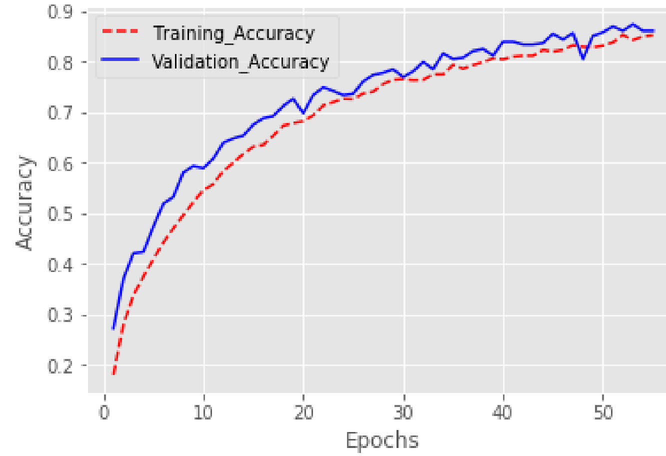
# Training/Validation

Loss: 0.4618090449643782



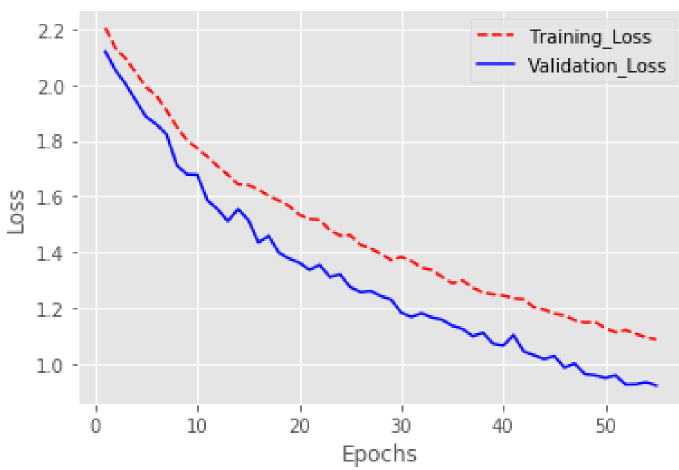

Loss MFC

Acc: 0.8371607499480994



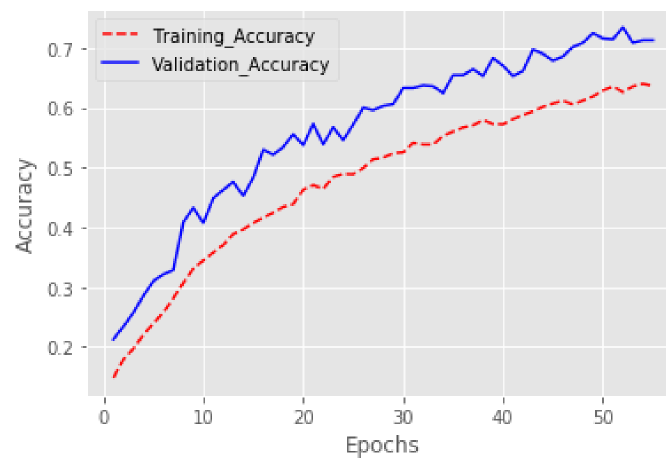

Accuracy MFC

Loss: 0.9861080977513547




Loss MFCCs

Acc: 0.7014613767506435



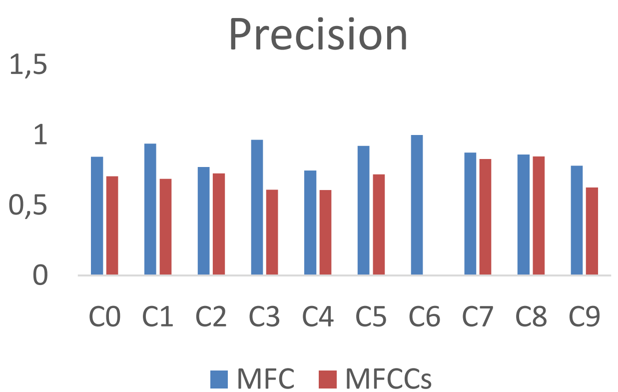

Accuracy MFCCs

# RESULTS MFC/MFCCs

| | Acc0 | Acc1 |
|---|---|---|
| MFC | 0.813 | 0.882 |
| MFCCs | 0.679 | 0.721 |

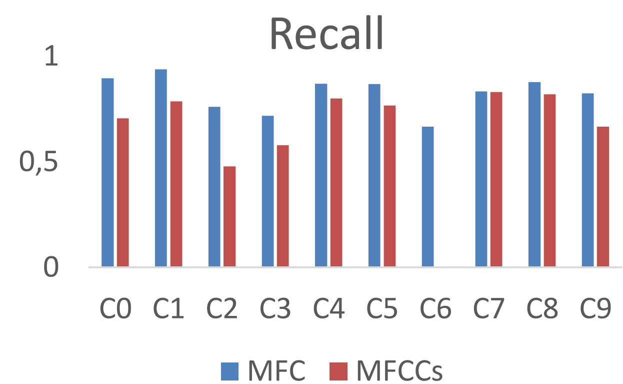
Acc0: Mean accuracy (not increased)

Acc1: Mean accuracy (increased)



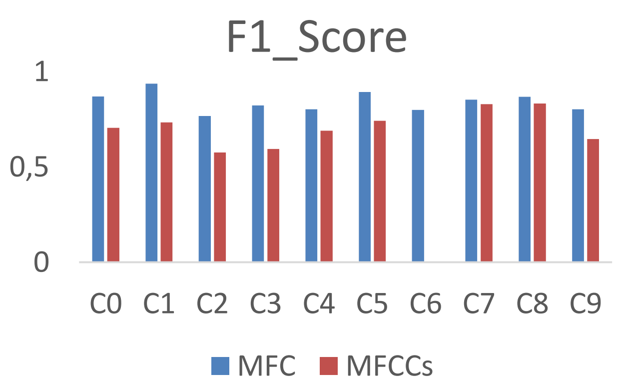

0.84507042, 0.9375 , 0.77272727, 0.96551724, 0.74603175,
0.921875 , 1. , 0.875 , 0.86 , 0.78205128

0.70588235, 0.6875 , 0.72727273, 0.61111111, 0.60759494,
0.71875 , 0. , 0.82978723, 0.84745763, 0.62650602




0.89552239, 0.9375 , 0.76119403, 0.71794872, 0.87037037,
0.86764706, 0.66666667, 0.83333333, 0.87755102, 0.82432432

0.70588235, 0.78571429, 0.47761194, 0.57894737, 0.8 ,
0.76666667, 0. , 0.82978723, 0.81967213, 0.66666667




0.86956522, 0.9375 , 0.76691729, 0.82352941, 0.8034188 ,
0.89393939, 0.8 , 0.85365854, 0.86868687, 0.80263158

0.70588235, 0.73333333, 0.57657658, 0.59459459, 0.69064748,
0.74193548, 0. , 0.82978723, 0.83333333, 0.64596273

Asiel Aldana Ortiz
<asiel.aldana89@gmail.com/>
<www.onclix.tech/>