An ESN approach for audio classification in construction sites

Edoardo Bini, Marco Ferraro, Alessandro Giannetti

OVERVIEW

INTRODUCTION

introduction to paper and state of the art

USE OF ESN

description of the approach used and the role of the ESN

IMPLEMENTATION

choices for the detection

RESULTS

results obtained

INTRODUCTION

MACHINERY AUDIO

Construction vehicles and tools

AUDIO SENSORS

environmental microphones, microphones placed on vehicles and inside them

AUDIO RECOGNITION

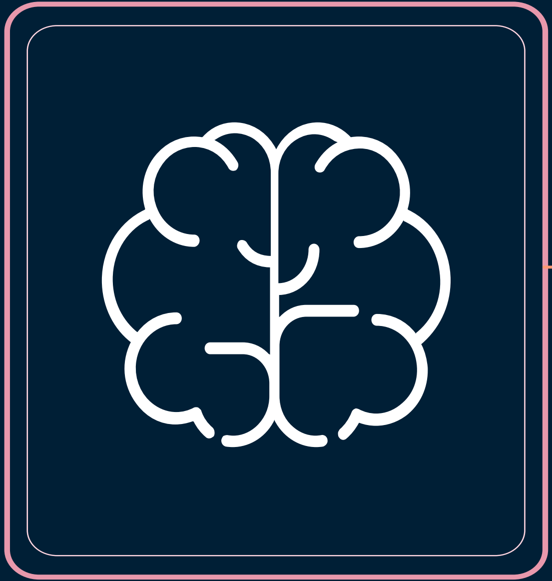
through an **E**cho **S**tate **N**etwork

PURPOSE OF THE PROJECT

INTRODUCTION
USE OF ESN
IMPLEMENTATION
RESULTS

Implement a Recurrent Neural Network (RNN) to classify active machinery in construction sites.

In detail an **Echo State Network**.



WHAT IS AN ESN?

ECHO STATE NETWORK

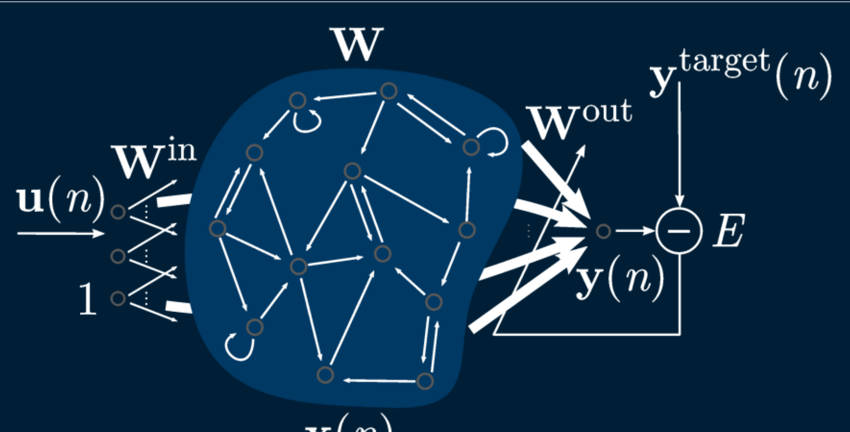
- Develop efficient learning algorithm solving RNN, signal processing and machine learning problems

- An important component are the **reservoirs**. It can serve as a memory, providing temporal context. Is defined as $(W^{in}, W,\ \alpha)$:
  - The input and the recurrent connection matrices are generated randomly according to some parameters

WHAT IS AN ESN?

RESERVOIR

- Defined as $(W^{in}, W,\ \alpha)$: The input and the recurrent connection matrices are generated randomly according to this parameters:
  - Distr. of nonzero elem.
  - Size N_x
  - Sparsity
  - Spectral radius W
  - Scaling of W^{in}
  - Leaking rate



Echo State Network

HOW DO ESNs WORK?

- Generate a Large random Reservoir RNN $(W^{in}, W,\ \alpha)$
- Run it using the training input and collect the reservoir activation states
- Compute the linear readout weights
- Use the trained network on new input data employing the trained output weights

DATASET

INTRODUCTION
USE OF ESN
IMPLEMENTATION
RESULTS

UTAH AUDIO DATA

- Audio collected from different construction machines and equipment from workers

- Real working scenarios – background noises

- Classes selected:
  - Backhoe JD50D Compact492
  - Compactor Ingersoll Rand
  - Concrete Mixer
  - Excavator Cat 320E
  - Excavator Hitachi 50U

DATASET

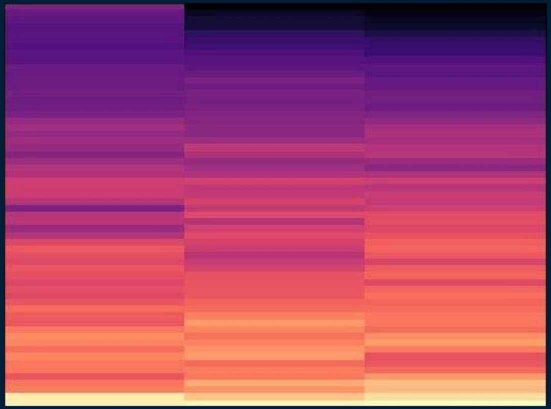
PREPROCESSING

- Split each audio file into 30 ms segments
- Discard segments under the average signal power by computing the Root Mean Square (mostly silent segments)

DATASET

PREPROCESSING

- Generation of log-scaled mel-spectrogram from the waveform of the audio tracks with a sampling of 44100 Hz [Librosa python library]



Example of mel-spectrogram

INPUT DATA ELABORATION

- Input of the ESN is a Numpy array concatenating the values of the three-time buckets of the spectrogram, one mel-band at a time

- The labels corresponding to each segment consist of the one-hot-encoding of the specific class, also in Numpy array

MEMORY USAGE ⚠

- EasyESN (Like most available ESN libraries) load the entire training dataset into memory before starting the training process
- Our training was forced into a **trade-off** between sizes of training data and the reservoir

POSSIBLE SOLUTION FOR FUTURE WORK

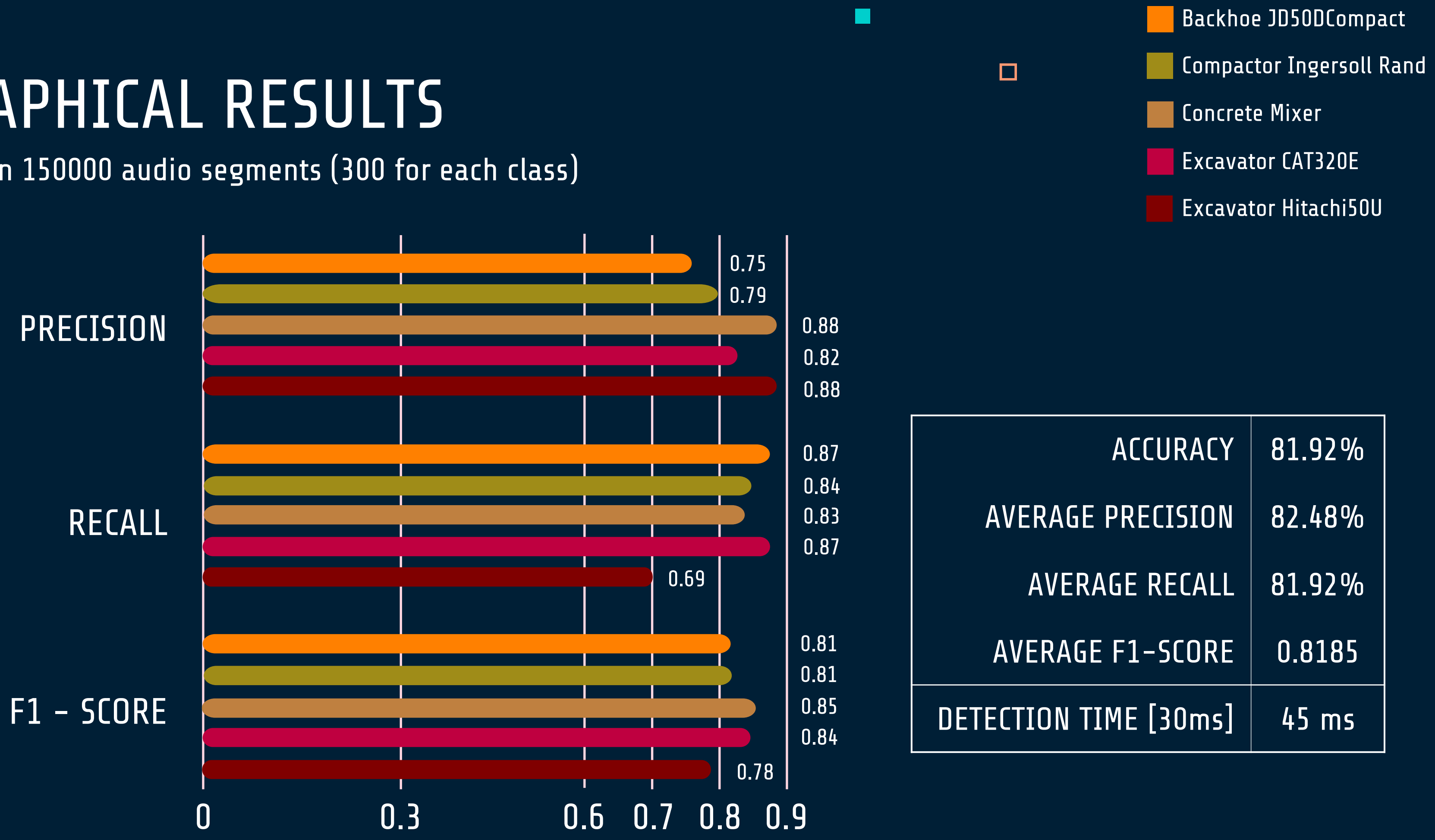
- Custom-made ESN with batch training
- Use dedicated large memory hardware

IMPLEMENTATIVE CHOICES

- **EasyESN** Python library
- Empirically found the best parameters for the ESN:
  - 300 audio segments for each class
    (1500 audio samples for the training set)
  - Reservoir size of 1200
  - Leaking rate of 0.1

GRAPHICAL RESULTS

Test on 150000 audio segments (300 for each class)



ACCURACY	81.92%
AVERAGE PRECISION	82.48%
AVERAGE RECALL	81.92%
AVERAGE F1-SCORE	0.8185
DETECTION TIME [30ms]	45 ms

CONFUSION MATRIX

	Backhoe JD50DCompact	Compactor Ingersoll Rand	Concrete Mixer	Excavator CAT320E	Excavator Hitachi50U
Backhoe JD50DCompact	26019	541	204	1157	2169
Compactor Ingersoll Rand	572	25119	1903	2214	192
Concrete Mixer	269	3217	24977	1376	134
Excavator CAT320E	1597	954	1185	26004	260
Excavator Hitachi50U	6047	2123	258	810	20762

POST-PROCESSING

MAJORITY VOTING

- Augment the capabilities of the classifier implementing a majority voting system

    - ~0,5 Seconds (17 segments)
    - ~1 Second (34 segments)
    - ~2 Seconds (67 segments)
    - ~3 Seconds (100 segments)

MAJORITY VOTING

CLASS PREDICTION

[0, 0, 0, 0, 0, 0, 3, 3, 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0,
3, 0, 0, 0, 0, 0, 3, 0, 4, 0, 4, 3, 4, 4, 0, 0, 0, 0, 0, ...]

~0,5 SECONDS (17 segments)

0, 0, 0, 0, 0, 0, 3, 3, 3, 4, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	...
0	0	...

~1 SECONDS (34 segments)

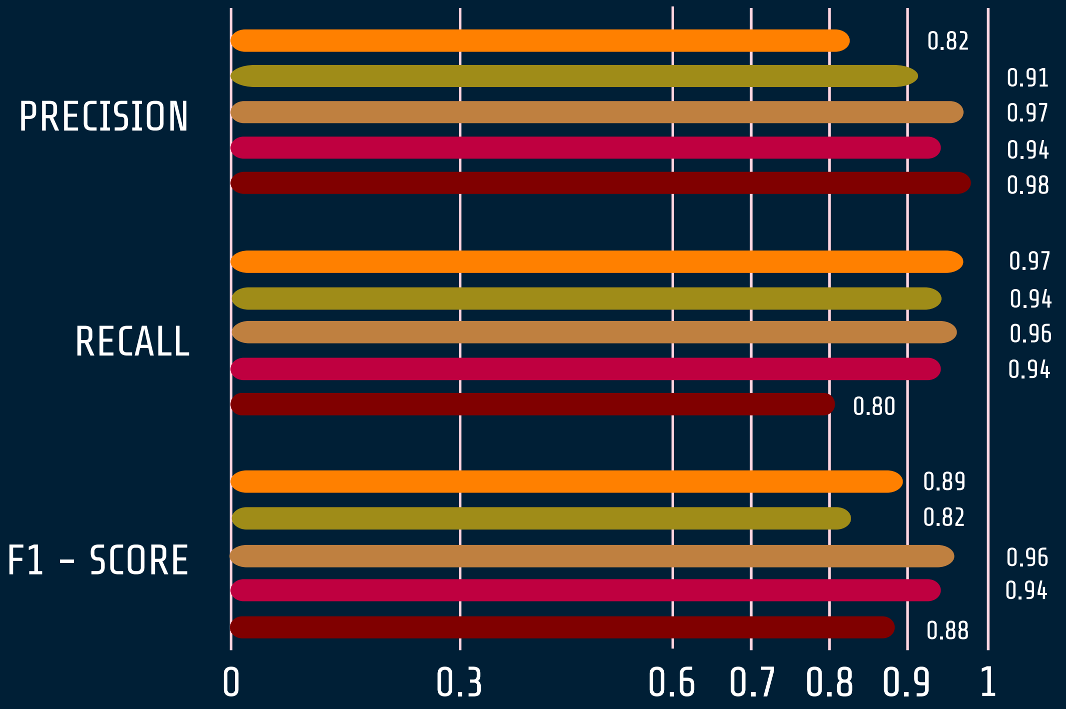
0, 0, 0, 0, 0, 0, 3, 3, 3, 4, 0, 0, 0, 0, 0, 0, 0 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	...
0	0	...

~2 SECONDS (67 segments)

0, 0, 0, 0, 0, 0, 3, 3, 3, 4, 0, 0, 0, 0, 0, 0, 0 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 3, 0 0, 0, 0, 0, 0, 3, 0, 4, 0, 4, 3, 4, 4, 0, 0, 0, 3 0, 0, 3, 3, 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0	...
0	0	...

MAJORITY VOTING

~0.5 SECOND (17 SEGMENTS)

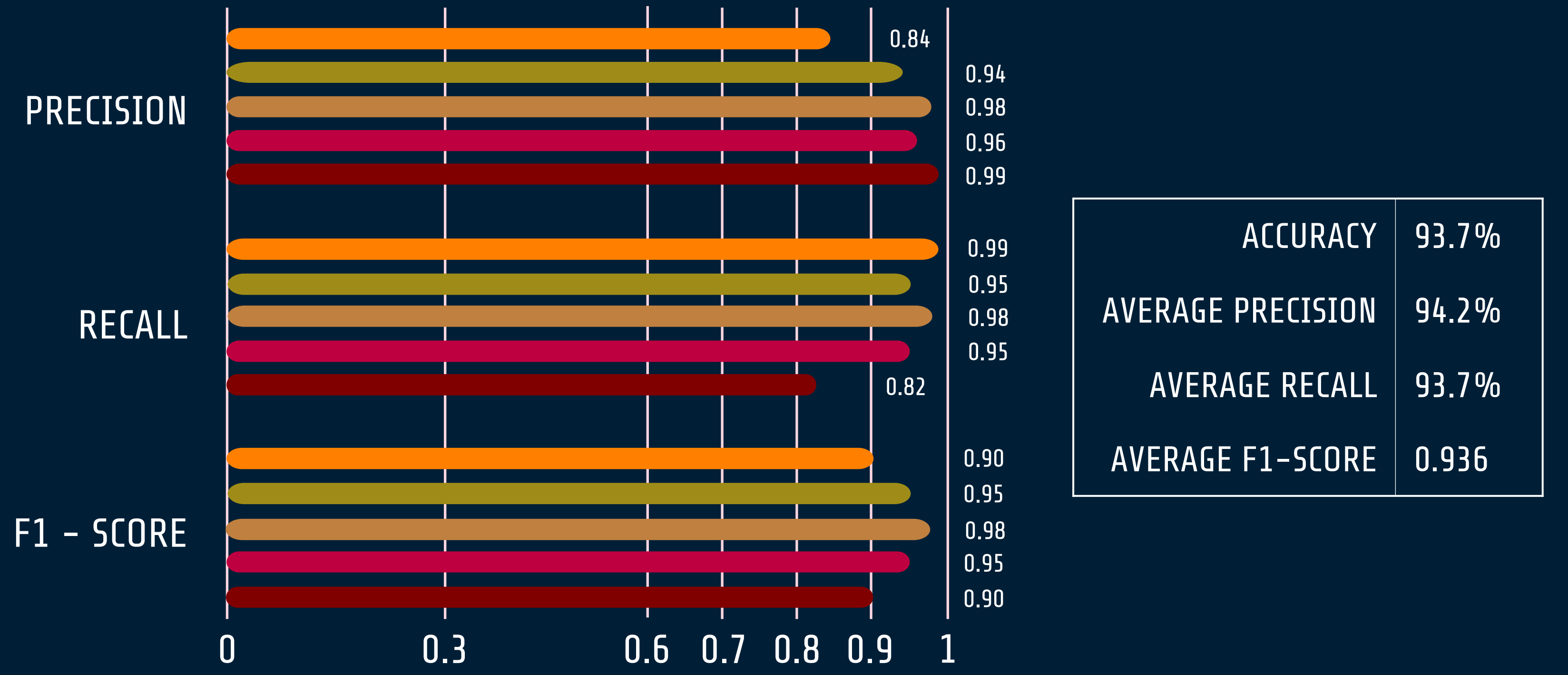


ACCURACY	91.9%
AVERAGE PRECISION	92.5%
AVERAGE RECALL	91.9%
AVERAGE F1-SCORE	0.918

MAJORITY VOTING

~1 SECOND (34 SEGMENTS)

INTRODUCTION
USE OF ESN
IMPLEMENTATION
RESULTS

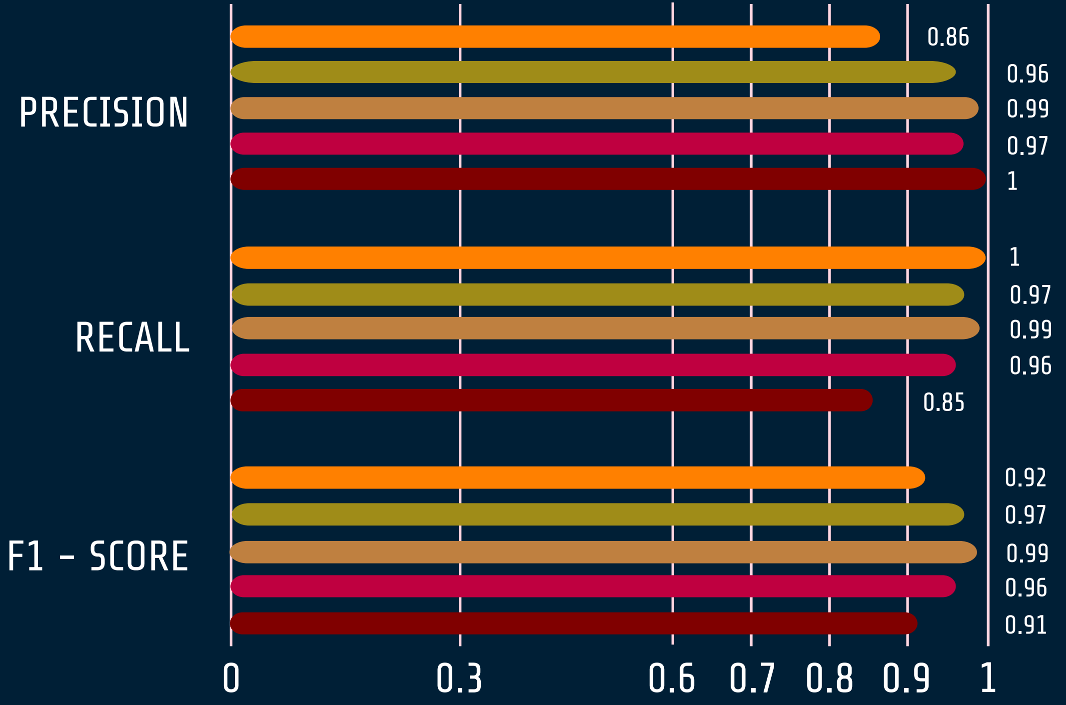


ACCURACY	93.7%
AVERAGE PRECISION	94.2%
AVERAGE RECALL	93.7%
AVERAGE F1-SCORE	0.936

MAJORITY VOTING

~2 SECOND (67 SEGMENTS)

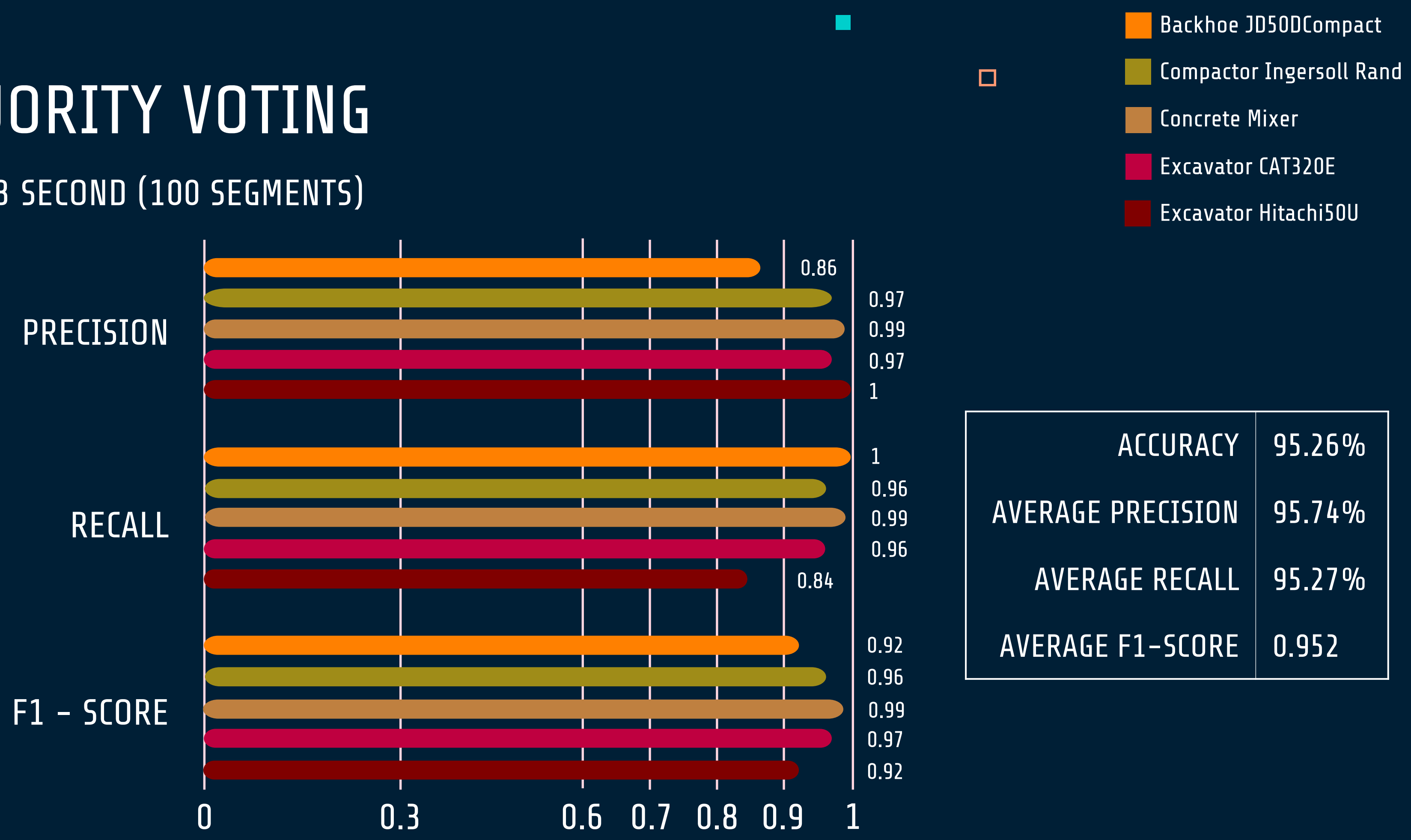
- Backhoe JD50DCompact
- Compactor Ingersoll Rand
- Concrete Mixer
- Excavator CAT320E
- Excavator Hitachi50U



ACCURACY	95.26%
AVERAGE PRECISION	95.72%
AVERAGE RECALL	95.26%
AVERAGE F1-SCORE	0.952

MAJORITY VOTING

3 SECOND (100 SEGMENTS)



ACCURACY	95.26%
AVERAGE PRECISION	95.74%
AVERAGE RECALL	95.27%
AVERAGE F1-SCORE	0.952

CONCLUSION

- ESNs demonstrated a remarkable versatility by showing their potential in the audio recognition field

- Unlike CNN it has worse performance, but it is easier to set up and much faster to train.

THANK YOU

For your attention