life.augmented

Sound Understanding and Image Classification Applications on STM32

Filippo Naccari – filippo.naccari@st.com

STMicroelectronics – SRA (System Research and Applications) – Catania Lab, IT.

May 22nd, 2026 – University of Catania – Department od Mathematics and Computer Science

Agenda

- # Toward Edge AI
- # Acoustic Scene Classification on STM32L4
- # Acoustic Event Detection + OOD on STM32H7
- # Image Classification + OOD on STM32H7

life.augmented

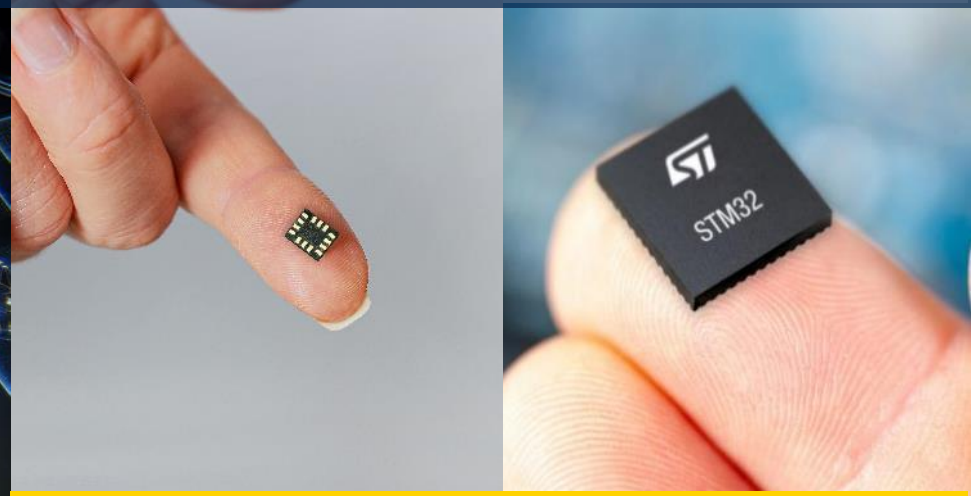
The explosion of AI-enabled devices is accelerating the inference shift from the cloud to the tiny edge

Inference **in the cloud** → Inference **on the device** → Inference **at the edge**

Cloud-centric

On device-centric

Tiny edge-centric

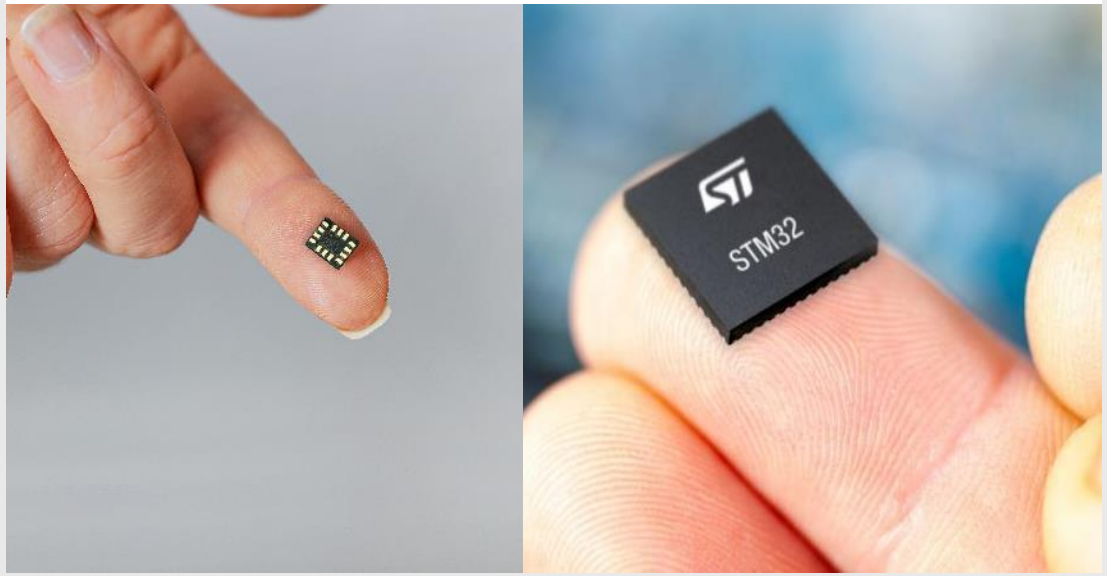


Enabled by a different class of hardware and software. Making AI more sustainable.

life.augmented

Inferring at the edge brings substantial benefits

01
10 **Reduced data transmission to generate meaningful information**



Ultralow latency
Real-time applications

Enhanced privacy and security
No data sharing in the cloud

Sustainable on energy
Low data, low power

Lower cost of inference to enable a new class of operations

life.augmented

Edge AI Applications

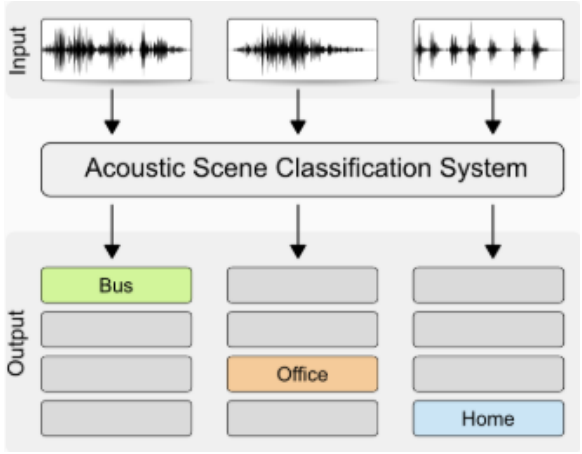
- **Edge AI** applications will be at the core of IoT nodes
- **IoT** node is mainly composed of:
  - Sensing (local sensors)
  - Computing (low power, ultra low power MCUs)
  - Connectivity (proximity, short range, long range)
- Edge AI Applications run near Sensors, where data are generated:
- Application complexity depends mainly on Sensor types:
  - Environmental Sensors (Temperature, Humidity, Pressure)
  - Inertial Sensors (Accelerometers, Gyroscopes, Magnetometers)
  - Acoustic Sensors (Digital MEMS Mics, Analogue UltraSound Mics, Mics Arrays)
  - Imaging/Video Sensors (Visible, IR, HDR,…)
  - 3D Sensors (ToF sensors, Lidar, Radar,…)

life.augmented

Acoustic Scene Classification on STM32L4

ASC (Task Description)

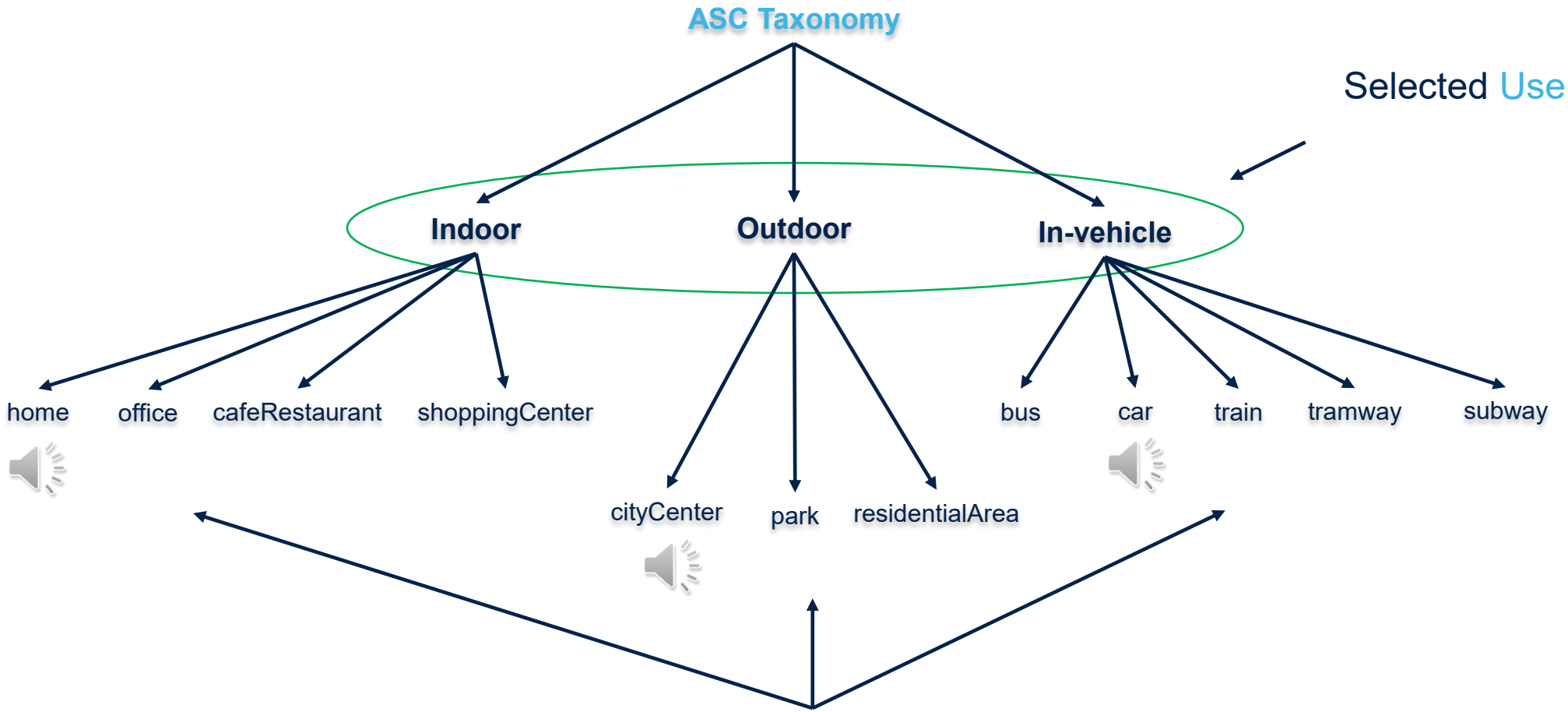
- **A**coustic **S**cene **C**lassification is the task of classifying environments from the sounds they produce



  - The starting point has been selected form DCASE (Detection and Classification of Acoustic Scenes and Events) challenges
  - The original task and dataset was 15 acoustic classes:
    - (beach, bus, cafe/restaurant, car, city_center, forest_path, grocery_store, home, library, metro_station, office, park, residential_area, train, tram)
  - A new dataset has been collected throughout an internal collection campaign

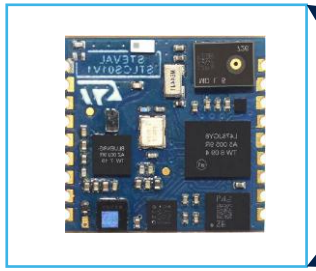
life.augmented

ASC Use Case and Taxonomy

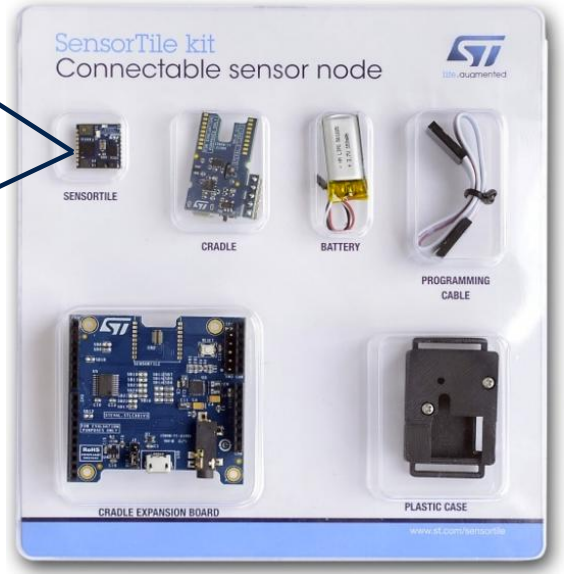


Target Platform - SensorTile

- An embedded version of the ASC has been implemented on ST target platform for a live demo application
- https://www.st.com/en/evaluation-tools/steval-stlkt01v1.html



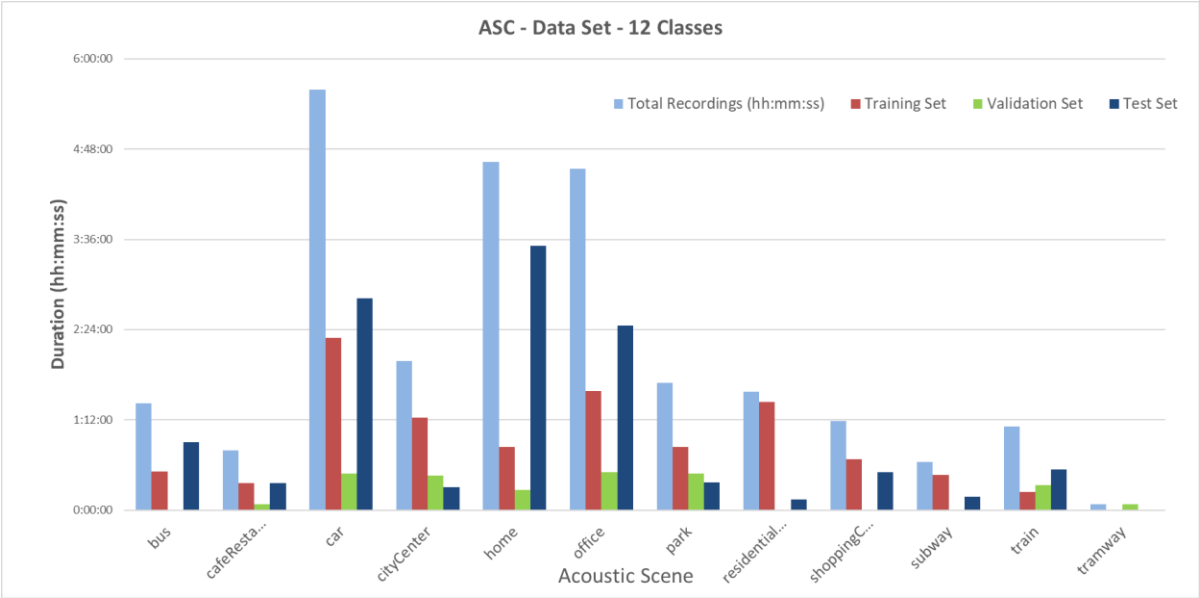
square shape
13.5 x 13.5 mm.

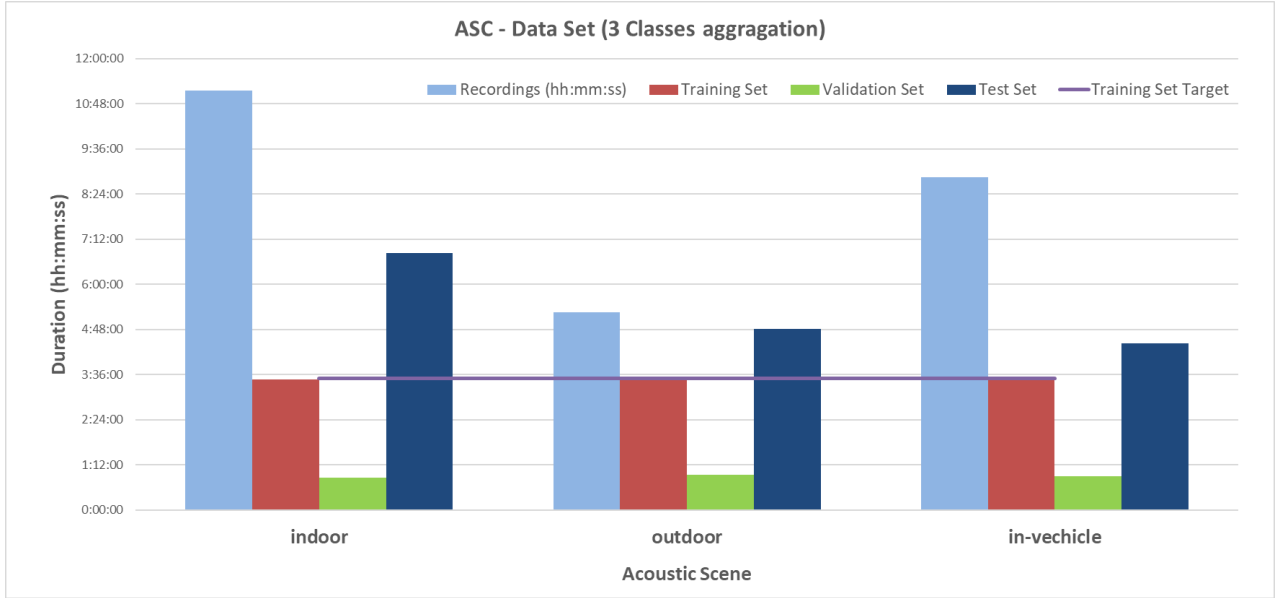


Main components:
– STM32L476JG – 32-bit ultra-low-power MCU with Cortex®M4F
– LSM6DSM – iNEMO inertial module: 3D accelerometer and 3D gyroscope
– LSM303AGR – Ultra-compact high-performance eCompass module: ultra-low power 3D accelerometer and 3D magnetometer
– LPS22HB – MEMS nano pressure sensor: 260-1260 hPa absolute digital output barometer
- HTS221 – capacitive digital sensor for relative humidity and temperature (STLCR01V1 cradle)
– MP34DT05-A – 64 dB SNR digital MEMS microphone
– BlueNRG-MS – Bluetooth low energy network processor

life.augmented

ASC - Dataset



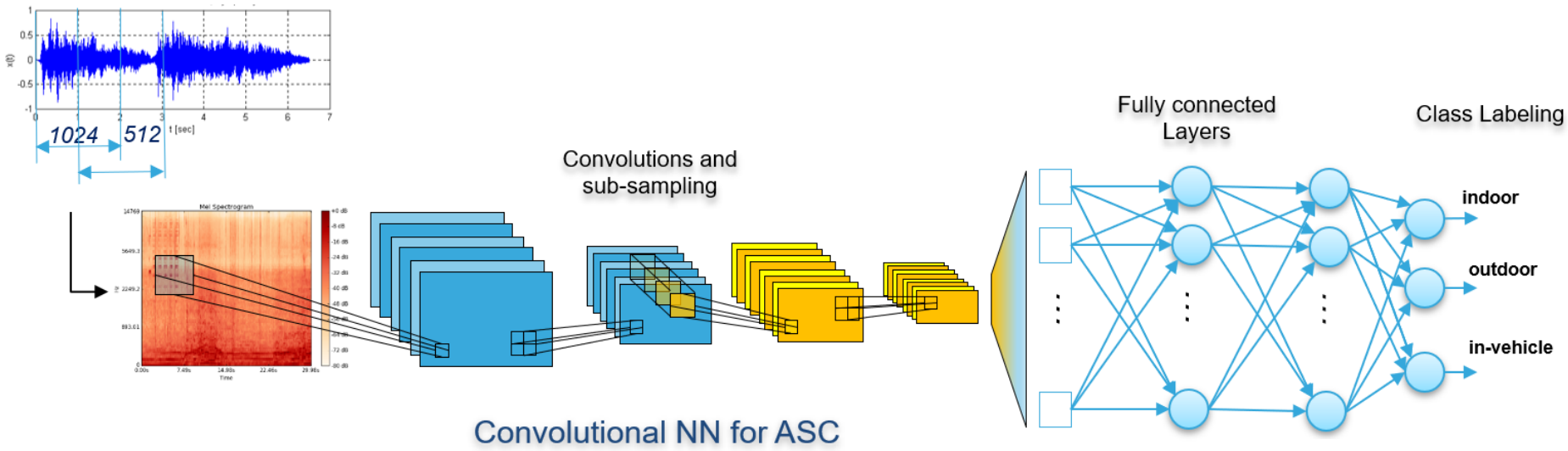


- Internal collected Dataset (~30h)
- Number of different recordings: ~150
- Audio PCM sampling rate: 16KHz (1ch)
- Resolution: 16 bps

Class	Training Set	Validation Set	Test Set
Indoor	3:28:01	0:51:17	6:49:37
Outdoor	3:30:23	0:56:32	4:48:23
In-vehicle	3:30:34	0:53:50	4:26:00
Total	10:28:58	2:41:39	16:04:00

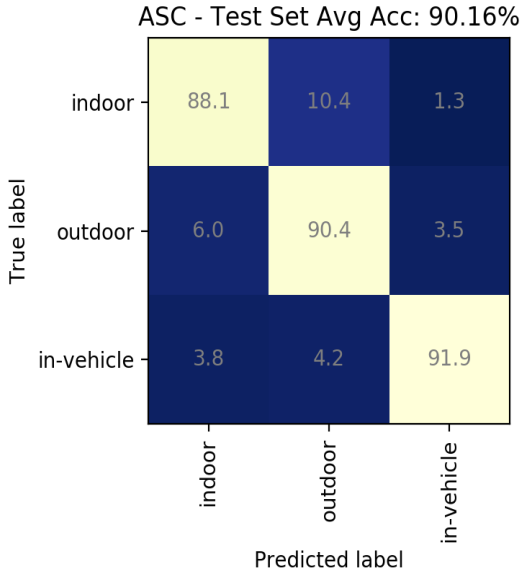
life.augmented

ASC – CNN Model



Convolutional NN for ASC

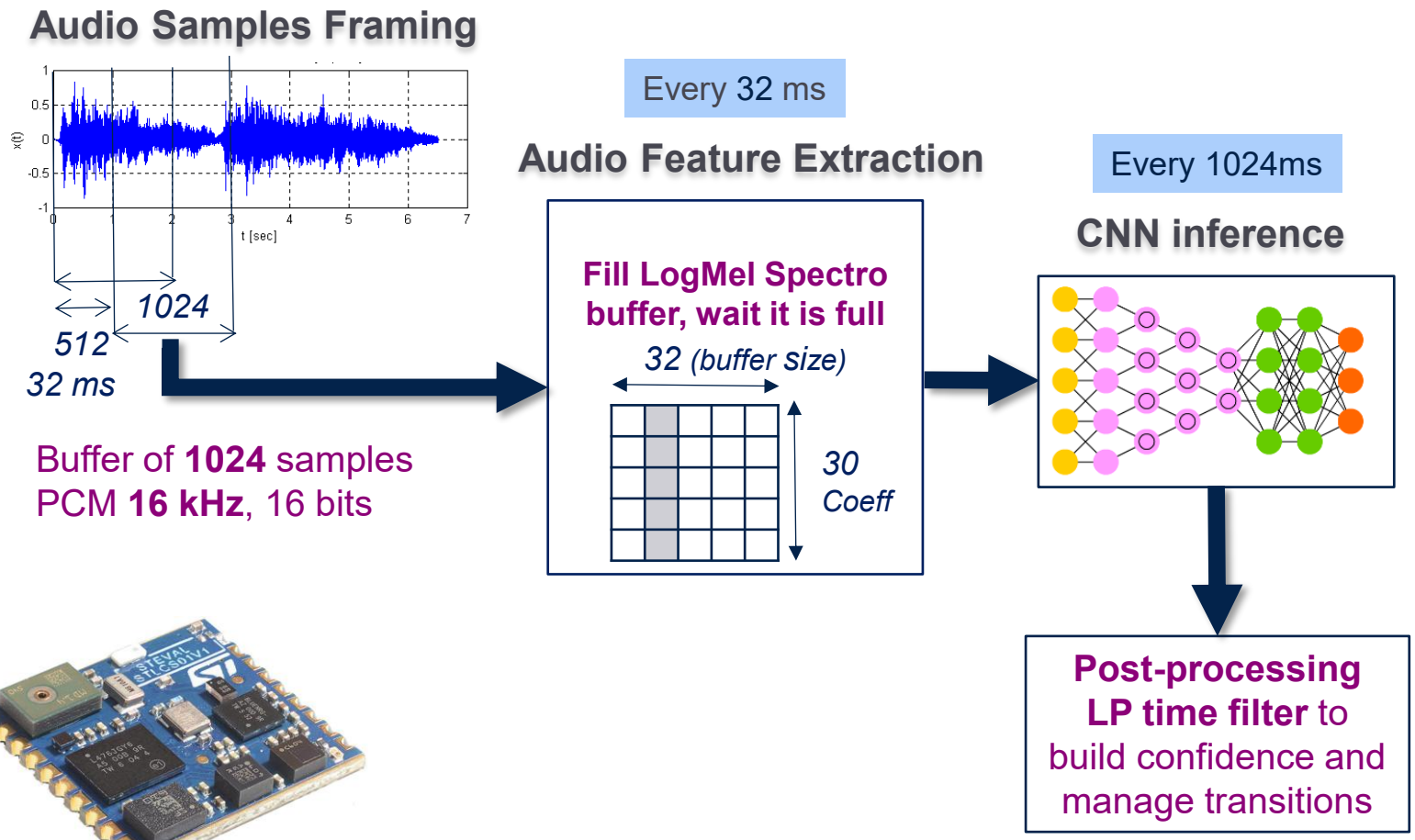
Layer	(type)	Out Shape	Param #
conv2d_1	(Conv2D)	(, 28, 30, 16)	160
max_pooling2d_1	(MaxPool2D)	(, 14, 15, 16)	0
batch_norm_1	(BatchNorm)	(, 14, 15, 16)	64
dropout_1	(Dropout)	(, 14, 15, 16)	0
conv2d_2	(Conv2D)	(, 12, 13, 16)	2320
max_pooling2d_2	(MaxPool2D)	(, 6, 6, 16)	0
batch_norm_2	(BatchNorm)	(, 6, 6, 16)	64
dropout_2	(Dropout)	(, 6, 6, 16)	0
flatten_1	(Flatten)	(, 576)	0
dense_1	(Dense)	(, 9)	5193
batch_norm_3	(BatchNorm)	(, 9)	36
activation_1	(Activation)	(, 9)	0
dropout_3	(Dropout)	(, 9)	0
dense_2	(Dense)	(, 3)	30
Total params			7867
Trainable params			7785
Non-trainable params			82



CNN model Confusion Matrix

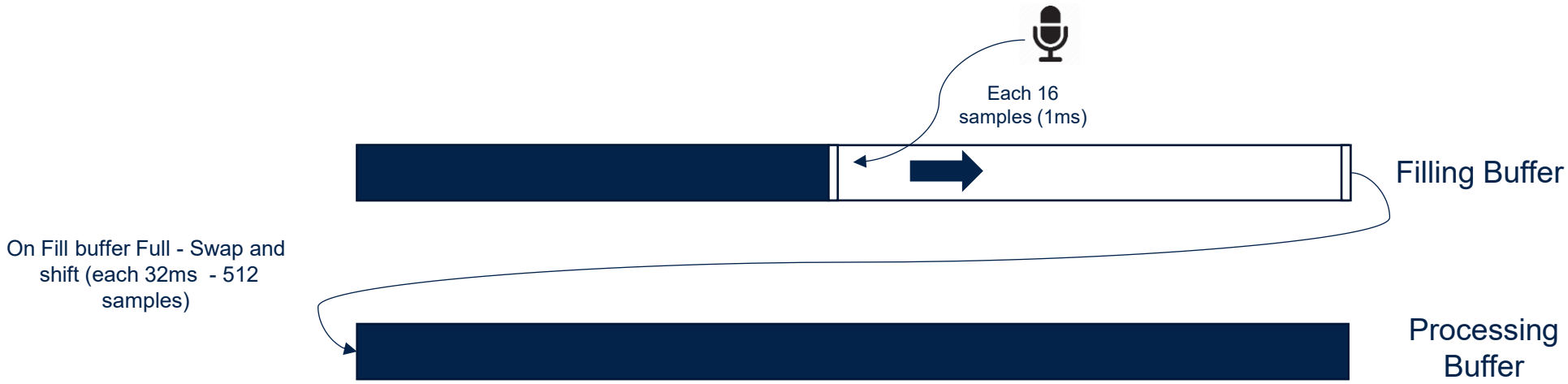
life.augmented

ASC on MCU


Audio Samples Framing
t [sec]
1024
512
32 ms
Buffer of **1024** samples
PCM **16 kHz**, 16 bits
Every 32 ms
Audio Feature Extraction
Fill LogMel Spectro buffer, wait it is full
32 (buffer size)
30 Coeff
Every 1024ms
CNN inference
Post-processing
LP time filter to build confidence and manage transitions

Audio Samples Buffering

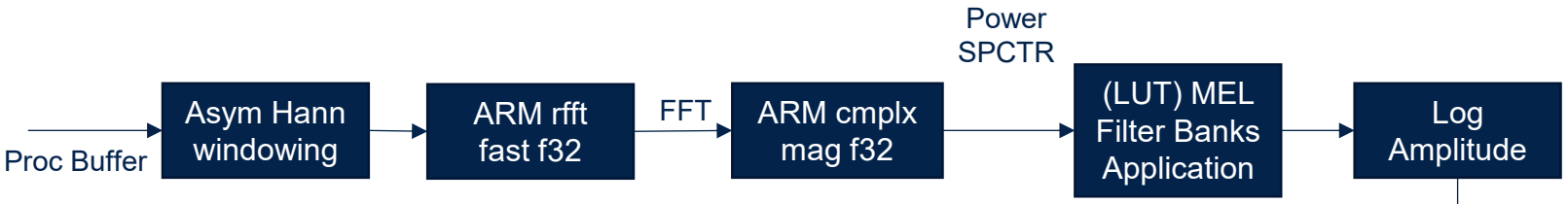
- Audio window buffer is composed of 1024 samples (64 ms @ 16KHz)
- Each new 512 samples (32 ms) the filling buffer is transferred to processing buffer and left shifted (audio frames are 50% overlapped)

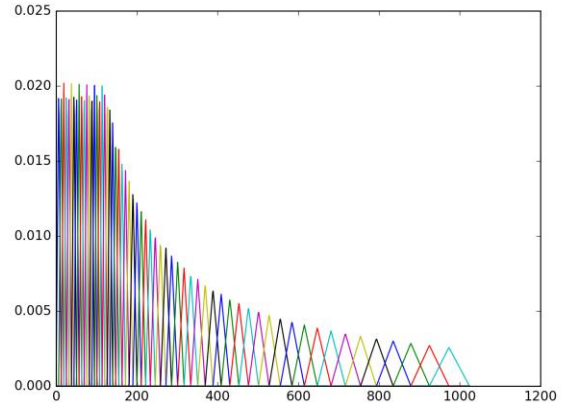


Filling Buffer (DMA driven) @1KHz : **540 cycles**
Processing buffer swap and shift @31.25Hz : **21.4 Kcycles**

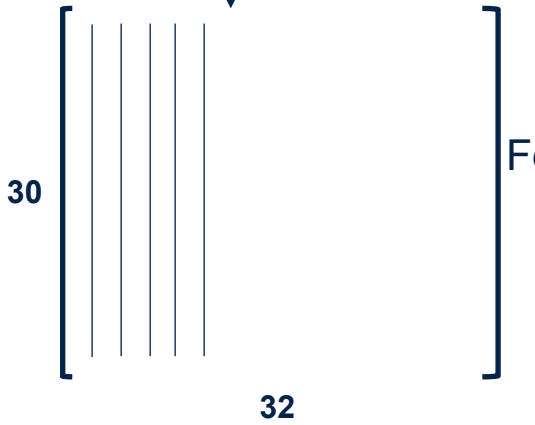
LogMel Energies Calculation

- Input: 1024 audio samples (64 ms)
- New frame feature calculation every 32ms (31.25Hz)
- Output: 30 Log Mel Energies (1 column of the 30x32 Feature Matrix)





Mel Filters Banks

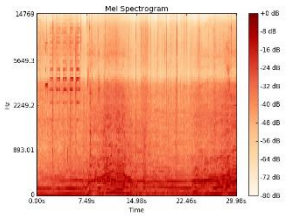


Feature Mtx: built on 1024 ms of audio samples

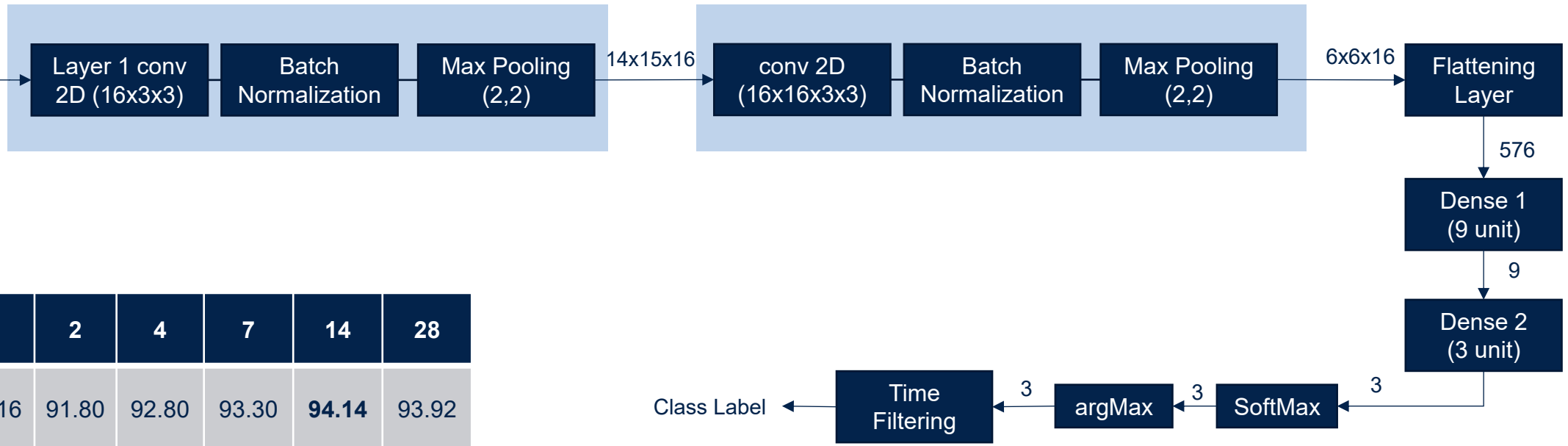
Execution time for one column calculation : **220 Kcycles** (2.75ms @80MHz on Cortex M4)

CNN 3CL Inference engine

- Input : 30x32 (960) LogMel Spectrogram Matrix
- Output: Class Label (0=indoor, 1=outdoor, 2=in-vehicle)
- X-CUBE-AI automatic NN mapping tool used to map Keras model into C code.
- PTQ (Post Training Quantization) has also been implemented from f32 to int8



Feature Mtx
(30x32)



Filter Length	-	2	4	7	14	28
Avg. Acc. (%)	90.16	91.80	92.80	93.30	**94.14**	93.92

Time Filtering Effects -
Accuracy vs Latency optimization

life.augmented

Complexity Figures on Cortex M4 MCU @80 MHz

Model	Avg. Acc. (%)	NVM (KB)	RAM (KB)	Inference Time (ms)
ASC CNN f32	90.16	30.81	17.41	81.505
ASC CNN int8	89.17	7.71	4.94	36.022

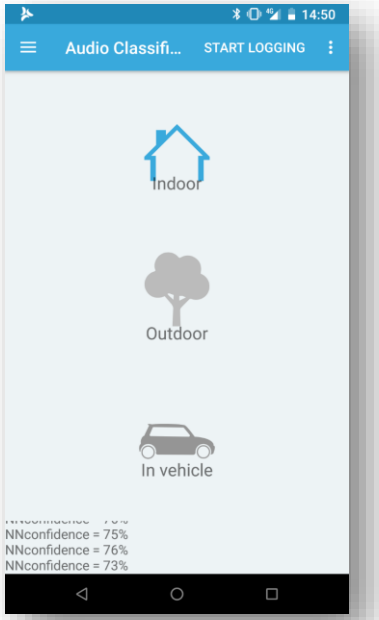
Inference Engine post training quantization to int8

Task	Task period (ms)	NVM (KB)	RAM (KB)	Exec. Time (ms)
PCM buffer managing	32	-	8.00	0.484
Feature Extraction	32	7.87	7.88	2.750
Inference Engine + PP	1024	7.71	4.94	36.072

Estimated MCU duty cycle of a **13.6%**

Resources

- Links
  - SensorTile development kit
    - https://www.st.com/en/evaluation-tools/steval-stlkt01v1.html
  - STM32Cube AI Studio
    - STEDGEAI-CUBEAI | Software - STMicroelectronics
  - Function Pack AI Sensing1
    - https://www.st.com/en/embedded-software/fp-ai-sensing1.html
  - Function Pack AI Context Awarness 1 (audio and motion sensing)
    - https://www.st.com/en/embedded-software/fp-ai-ctxaware1.html
  - ST BLE Sensor (Mobile APP)
  - https://play.google.com/store/apps/details?id=com.st.bluems&hl=en (Android)
  - https://apps.apple.com/it/app/st-ble-sensor/id993670214 (iOS)



Audio

Audio + Motion

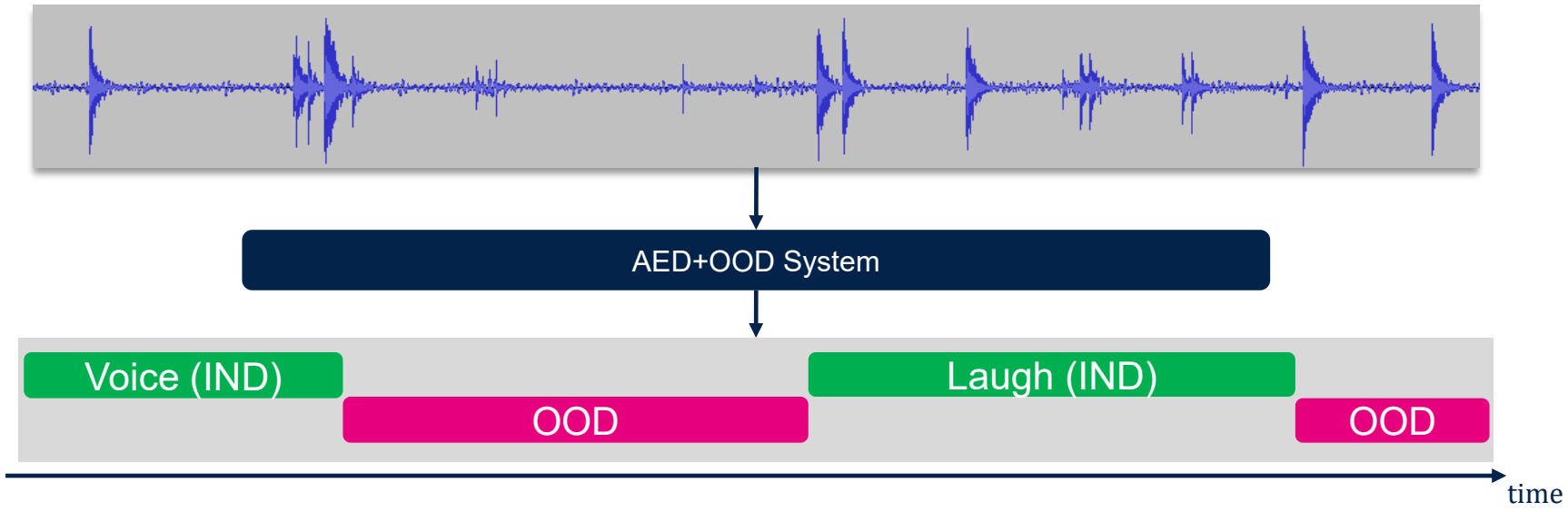
Acoustic Event Detection + OOD on STM32H7

Introduction

- **AED (Acoustic Event Detection)** is the task of recognizing acoustic events from an audio stream in sound understanding applications, e.g. :
  - Human sounds, Animal sounds, Music, Natural sounds, Sounds of Things, etc…

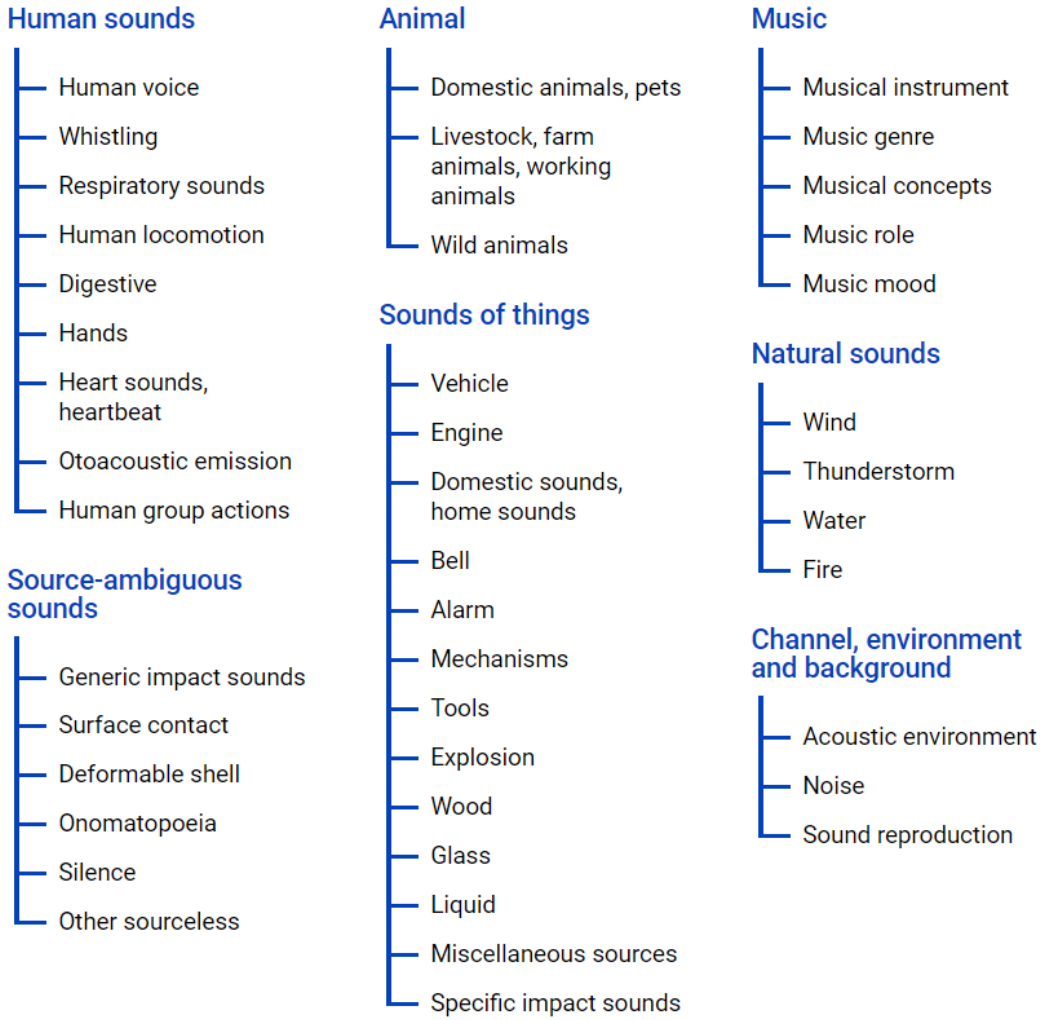
- **OOD (Out of Distribution) Detection** is the ability to detect samples that belong to unknown classes, not seen during training.

- **AED+OOD** example: recognize only human sounds and reject all other sounds.



YAMNet and AudioSet (Transfer Learning)

- **YAMNet** is a pre-trained deep net that predicts **521 audio event classes** based on the AudioSet-YouTube corpus and employing the Mobilenet_v1 depthwise-separable convolution architecture.
- **AudioSet** consists of an ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos.
  - The ontology is specified as a hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds.



AudioSet Taxonomy

life.augmented

AED Modeling Dataset

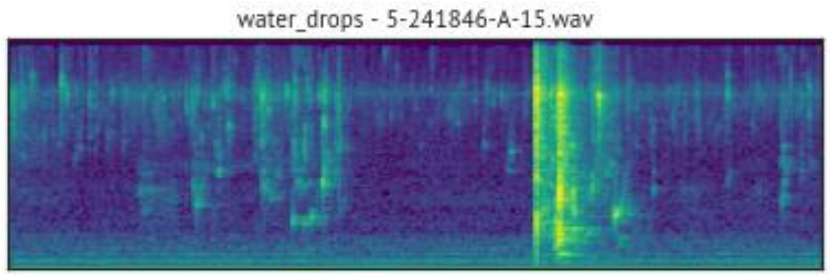
- A Data Augmented AED dataset based on ESC-50 public dataset:
  - ESC-50 main features: 2.000 samples, 50 classes, 40 samples/class, each of 5 secs.
- Data Augmentation process based on synthetic mixture process using:
  - Automatic Sequence trimming based on Signal Activation Detector (SAD)
  - Time stretch in a realistic range
  - Reverberation
- DA Modeling dataset features:
  - 50.500 samples
  - Samples length 1 sec
  - Augmentation factor 25.25x

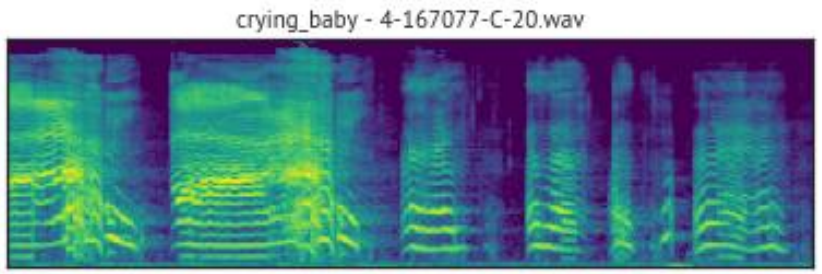
Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

life.augmented

PCM Samples Framing and Featurization

- PCM samples are framed and featurized according to YAMNet Encoder Input requirements:
  - PCM samples @ 16 kHz mono.
  - A Spectrogram is computed using magnitudes of the Short-Time Fourier Transform with a window size of 25 ms (400 samples), a window hop of 10 ms (160 samples).
  - A Mel Spectrogram is computed by mapping the spectrogram to 64 Mel bins covering the range 125-7500 Hz.
  - These features are then framed into 50% overlapping windows of 960 ms, leading to a 96x64 input to the Feature Encoder Net.



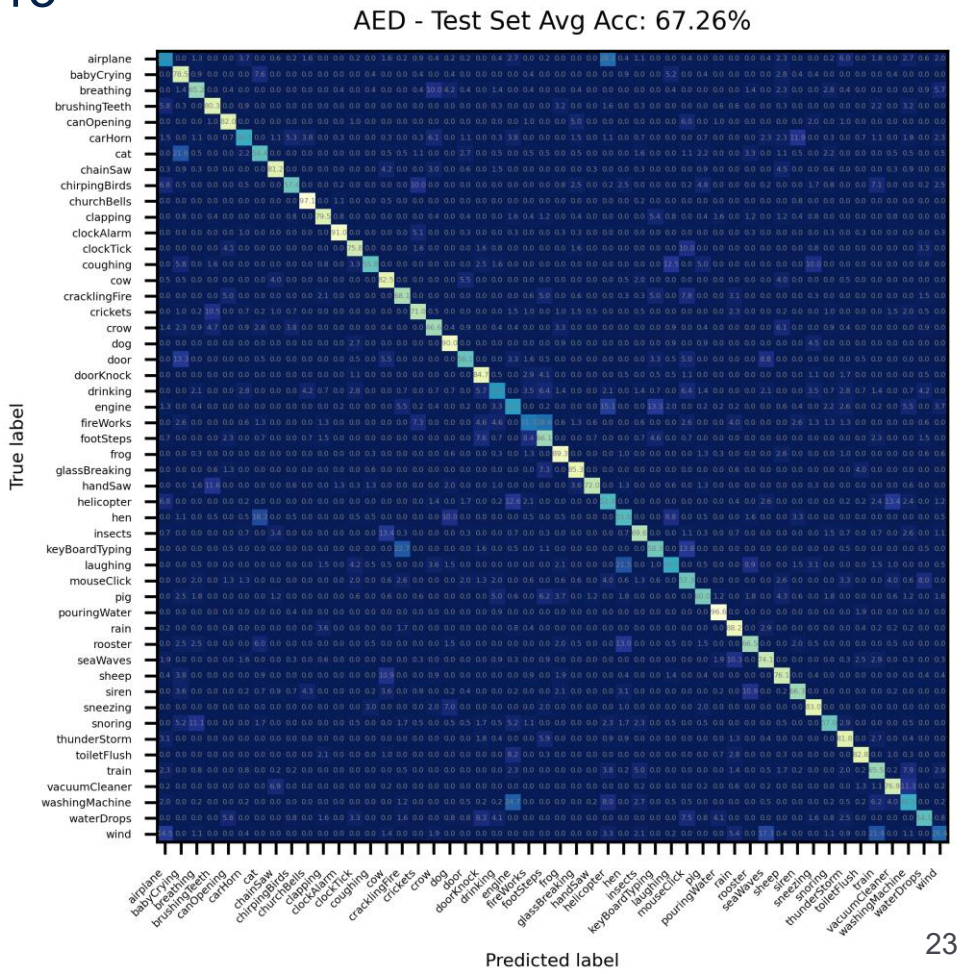


life.augmented

Pre Trained Yamnet versions on DA13 (ESC-50)

- Tested AED modeling in Transfer Learning mode on different pre-trained YAMNet f32 versions for the 50 classes DA13 (Esc-50 augmented version) dataset
  - Top 1 accuracy on single spectrogram on 50 classes of DA13

Model	Input Size	# Classes	Avg. Acc. (%)	TL #Trainables	Yamnet #Trainables
Yamnet cut @ 128	96x64	50	51.58	6,450	32,192
Yamnet cut @ 256	**96x64**	**50**	**67.26**	**39,858**	**138,176**
Yamnet cut @ 512	96x64	50	46.43	25,650	1,626,048
Yamnet 1024 (baseline)	96x64	50	60.37	51,250	3,217,344



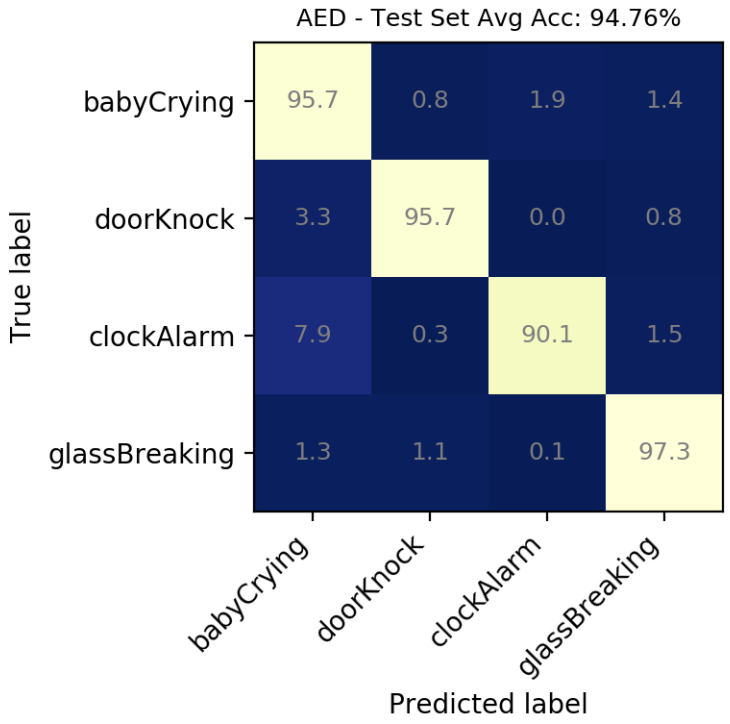
life.augmented

Use Cases on Yamnet 256 embeddings

- More flexibility added to use case definition inside the perimeter of ESC-50 dataset.
- Different TL models (f32) have been generated from Yamnet 256 (f32) embeddings
  - **Mix1 (Home Sounds - 10 classes):**
    - 'babyCrying', 'clockAlarm', 'clockTick', 'coughing', 'dog', 'doorKnock', 'glassBreaking', 'laughing', 'vacuumCleaner', 'washingMachine'
  - Mix2 (Urban Sounds - 11 classes):
    - 'airplane', 'carHorn', 'churchBells', 'engine', 'fireWorks', 'helicopter', 'rain', 'siren', 'thunderStorm', 'train', 'wind'
  - Mix3 (Human not Speech Sounds - 10 classes):
    - 'babyCrying', 'sneezing', 'clapping', 'breathing', 'coughing', 'footSteps', 'laughing', 'brushingTeeth', 'snoring', 'drinking'
  - Mix4 (Natural and water Sounds - 10 classes):
    - 'rain', 'seaWaves', 'cracklingFire', 'crickets', 'chirpingBirds', 'waterDrops', 'wind', 'pouringWater', 'toiletFlush', 'thunderStorm'
  - **Mix5 (Home Alert Sounds - 4 classes):**
    - 'babyCrying', 'doorKnock', 'clockAlarm', 'glassBreaking'
  - Mix6 (Glass breaking - 1 class): 'glassBreaking'
  - Mix7 (Baby Crying - 1 class): 'babyCrying'

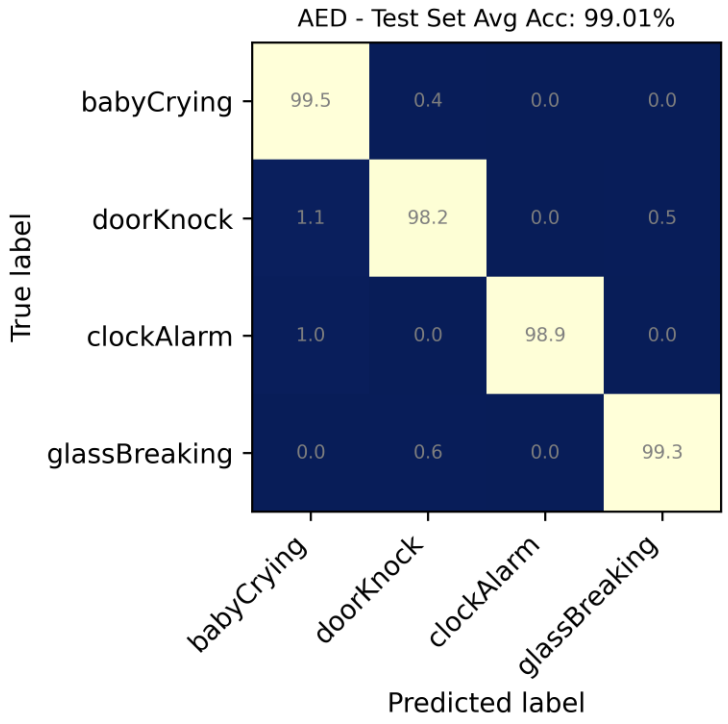
Home Alert Sounds use case (TL improvement)

AED Model on Spectrograms



- Input : 30x14 LogMel Spectrograms
- Model size: 9,848 trainables

AED Model on Yamnet 256 embeddings (TL)

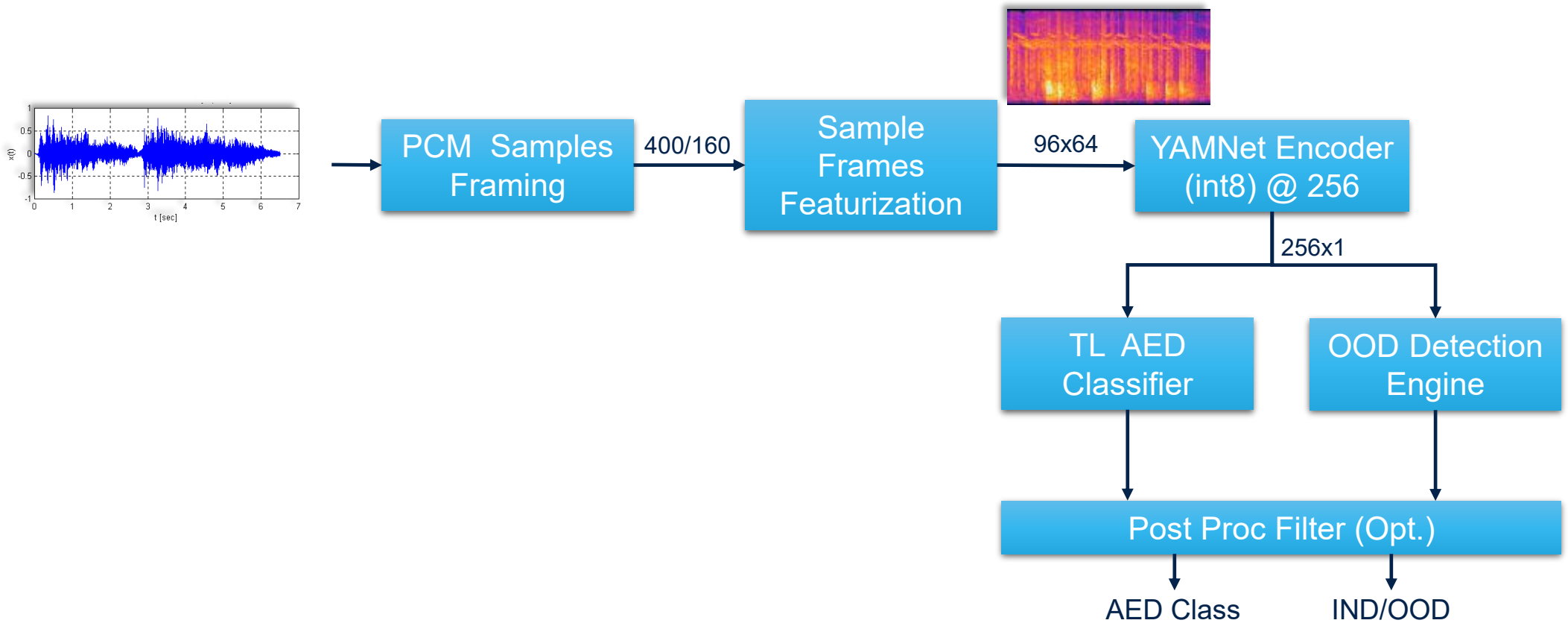


- Input : 96x64 LogMel Spectrograms
- Model Size: 1,028 + 138,176 Parameters

life.augmented

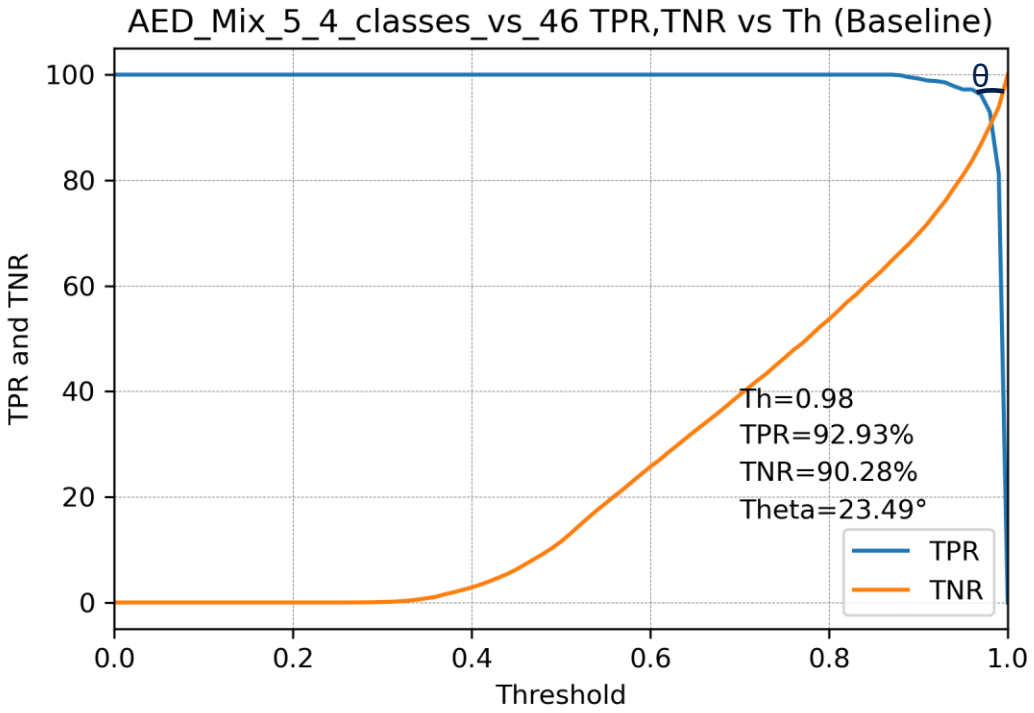
AED+OOD on Edge Platform

- Key Features of the proposed system for Acoustic Event Detection with OOD detection:
  - Quantized Feature Encoder (YAMNet) to enable Transfer Learning
  - Multiple TL Inference Engines for AED and OOD detection

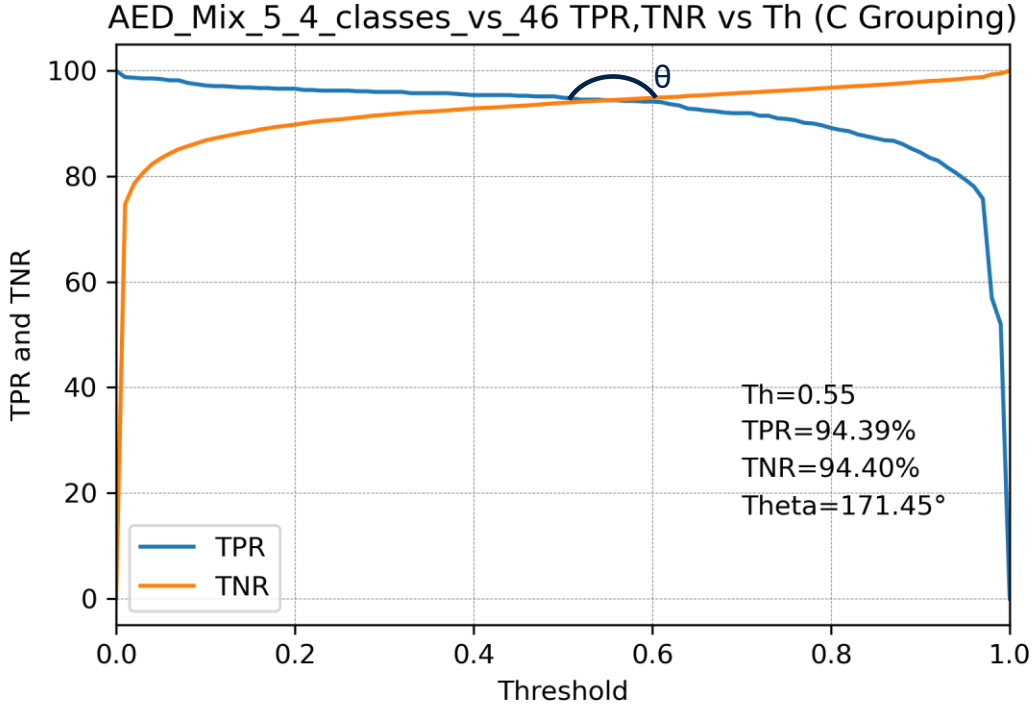


OOD on Home Alert Sounds (4 classes)

IND Classes: *Baby Crying, Door Knock, Glass Breaking, Clock Alarm*



Baseline with optimized Threshold
IND vs OOD Detection Performance



OOD Developed Solution Detection Performance
+ 2.8% Avg Bal. Accuracy
+ 147.96 deg on Theta

life.augmented

Inference Engines Profiling (4 classes)

- Use case: home alert sounds 4 classes (*babyCrying, glassBreaking, doorKnock , clockAlarm*)
- Complexity figures on STM32H7 @ 480MHz
- Task period: 480ms

Inf. Engine	Input size	# Trainables	NVM (KB)	RAM (KB)	MACCs	Inference Time (ms)
Yamnet 256 int8	96x64	134,720	136.62	124.88	23,976,128	77.5
AED TL (f32)	256x1	1,028	4.02	1.03	1,088	0.013
OOD (f32)	256x1	1,291	5.04	1.03	1,306	0.016
Total			**145.68**	**126.94**	**23,978,522**	**77.529**

life.augmented

Fruit Recognition + OOD on STM32H7

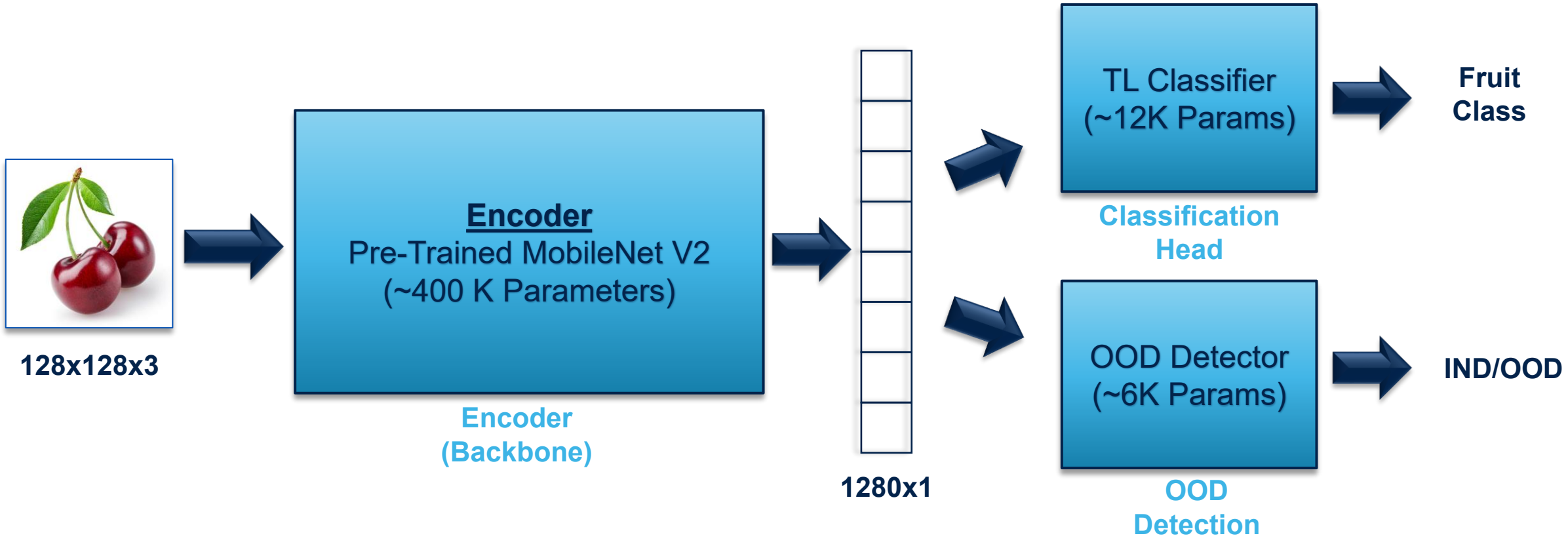
Fruit Recognition + OOD Use Case

- Transfer Learning from MobileNetV2
  - pre-trained on large scale dataset
  - (input = 128x128x3, alpha = 0.35)
- Training Dataset from Fruits-262 public dataset
  - https://www.kaggle.com/datasets/aelchimminut/fruits262
    - Number of classes: 262 fruits.
    - Total number of images: 225,640.
    - Number of images per class on average: 861

- Selected 10 classes of common fruits. (easily reconfigurable with alternative or more classes)
- Added Out Of Distribution Detection Model
  - It labels as **“unknown”** images not belonging to target classes.

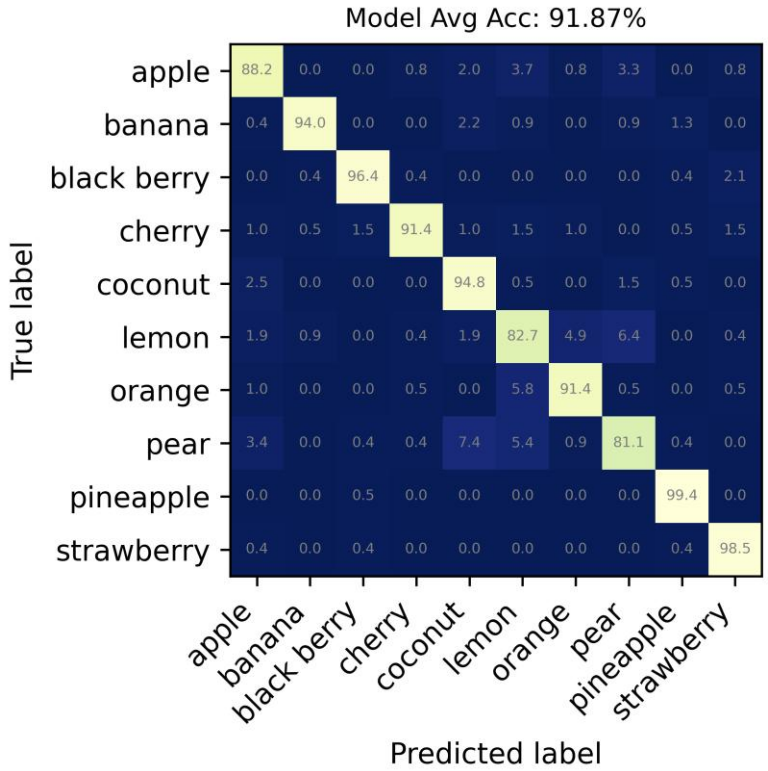
Models Set

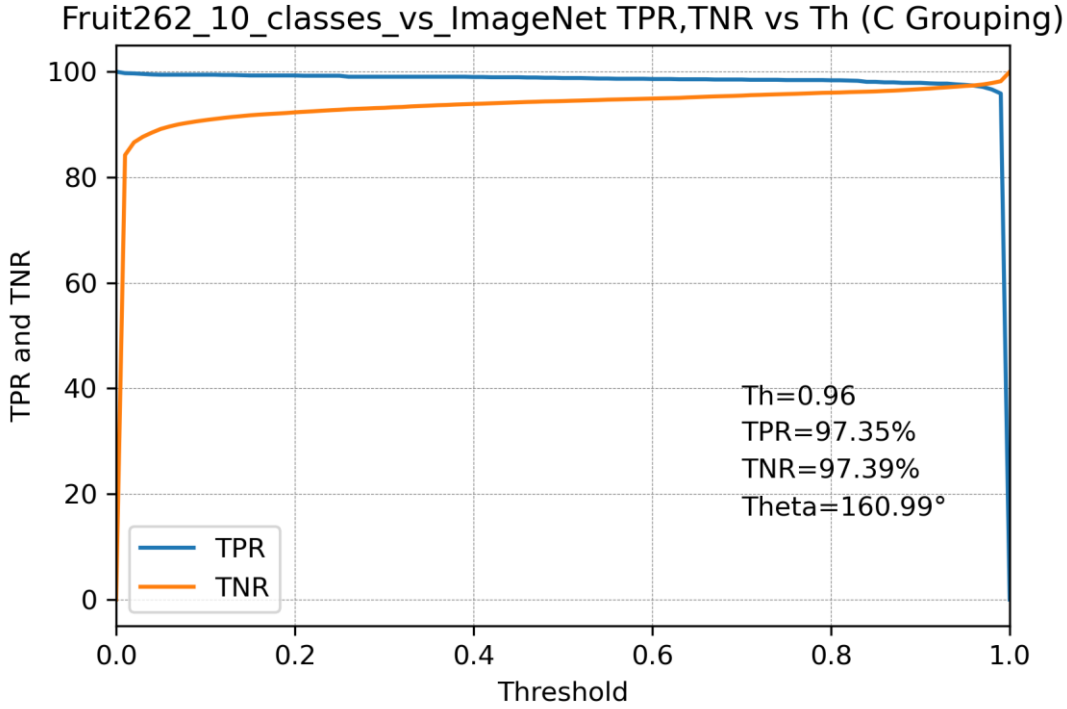
Fruit Recognition + OOD Multiple Model Setup



Fruits Classification + OOD Accuracy

- Classification Accuracy and OOD Detection performance:
  - Average Classification Accuracy: **91.87%**
  - OOD Detection Balanced Accuracy: **97.37% (+27.8% wrt baseline)**





life.augmented

Models Set Profiling

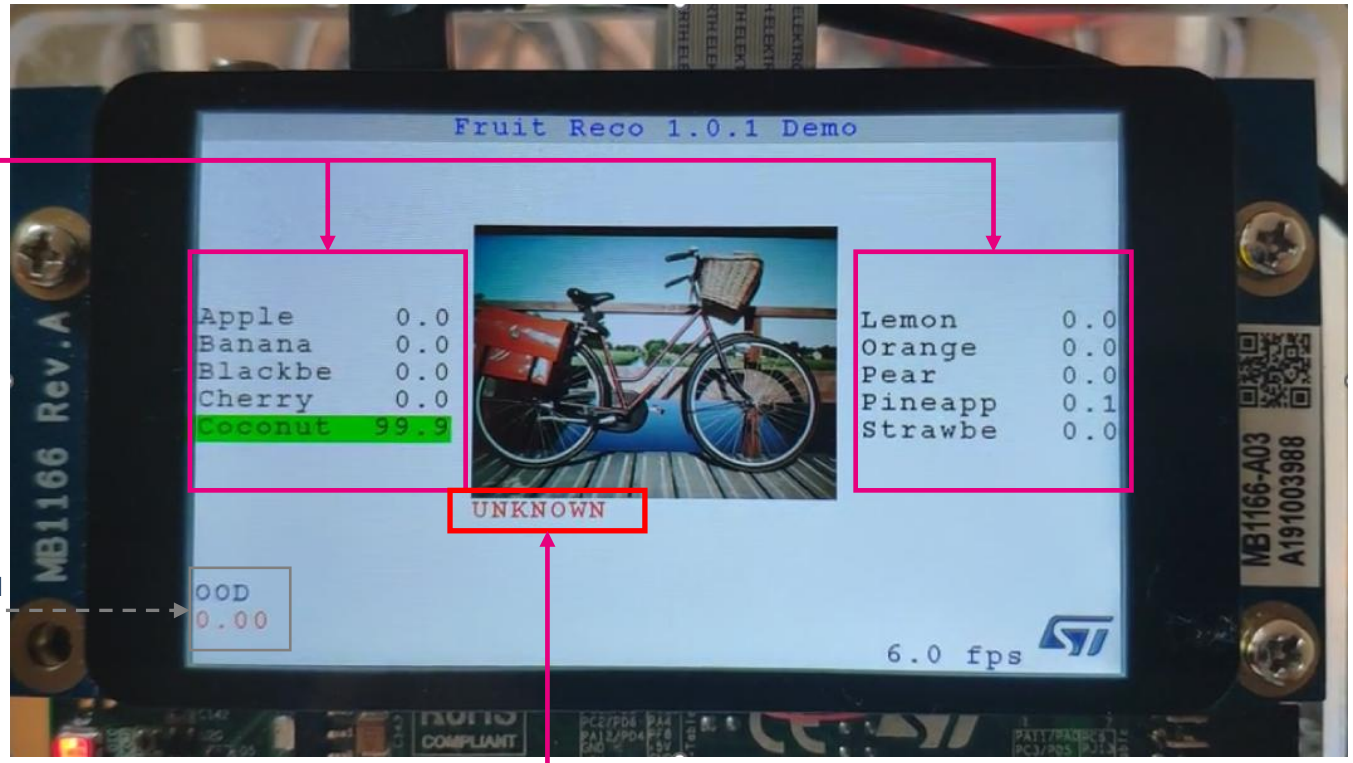
- Fruit Classification + OOD Application
  - Models complexity on STM32H747I-DISCO
  - **STM32Cube.AI Developer Cloud 8.1.0**

Inference Engine	Input size (I/O type)	# Params	NVM	RAM	MACCs	Inference Time (ms)
Encoder (f32) (MobileNetV2 128 0.35)	*128x128x3 (f32/f32)*	*396,128*	*1.5MB*	*824KB*	*20,687,744*	*514.3*
Encoder (int8) (MobileNetV2 128 0.35)	*128x128x3 (uint8/f32)*	*389,090*	*511KB **(-66.7%)***	*257KB **(-68.8%)***	*19,194,240*	***100.8 (>5x)***
TL Classifier (f32)	*1280x1 (f32/f32)*	*12,810*	*59 KB*	*6 KB*	*12,960*	*0.1827*
OOD Detector (f32)	*1280x1 (f32/f32)*	*6,411*	*35 KB*	*6 KB*	*6,426*	*0.094*
Total			**605 KB**	**269 KB**	**19,213,626**	**101.08**

life.augmented

Fruits Classification + OOD Detection Demo

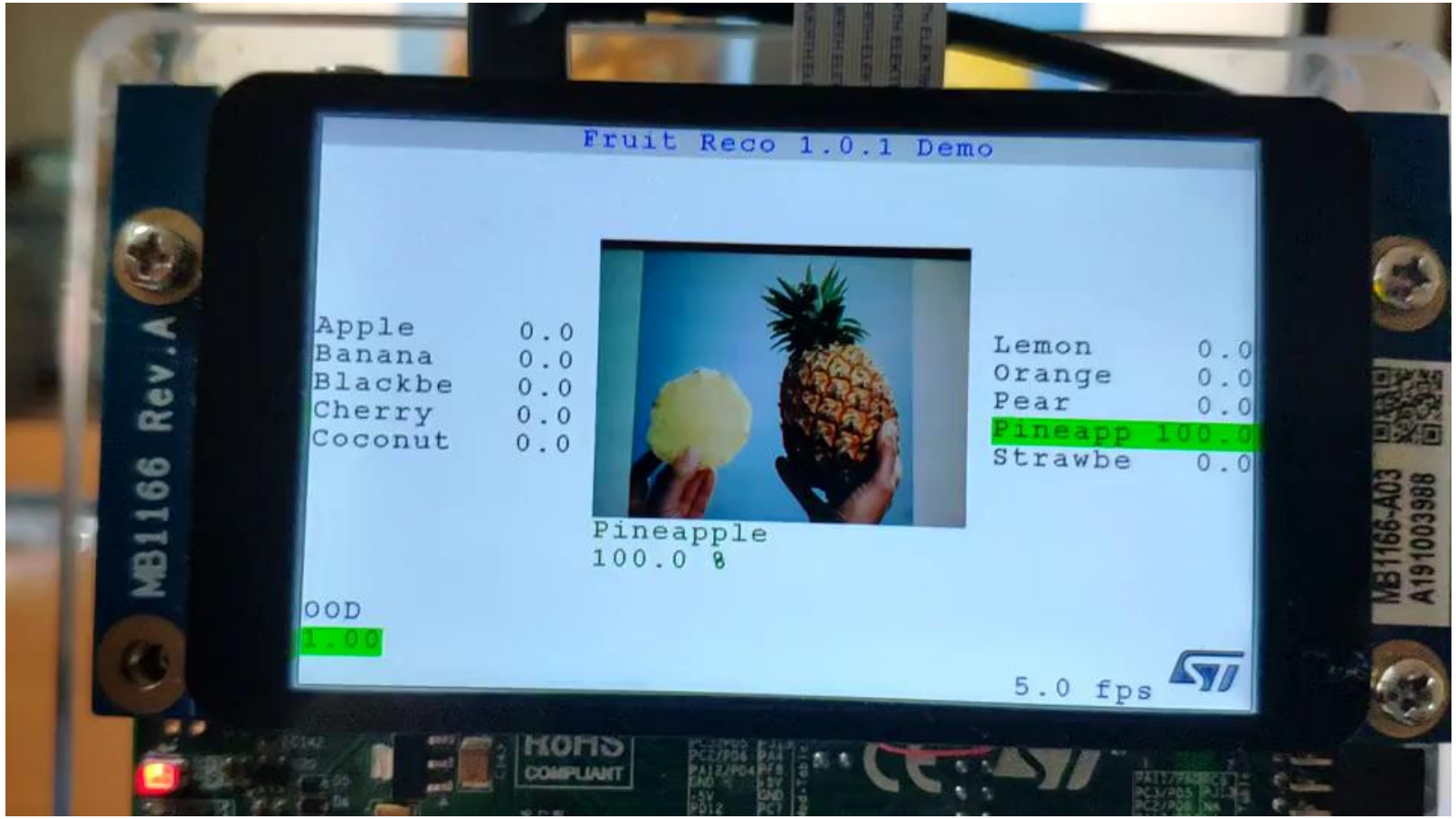
- Original model : Fruit Recognition (10 classes)
  - Trained in Transfer Learning on MobileNetV2, image resolution 128x128, alpha=0.35
- Added OOD Detection support
- Decision pipeline integrated in STM32H747I-DISCO



life.augmented

Demo Video on STM32H7

- Fruit Recognition + OOD Detection on STM32H747I-DISCO – ver 1.0.1



life.augmented

Our technology starts with You

Find out more at www.st.com

life.augmented