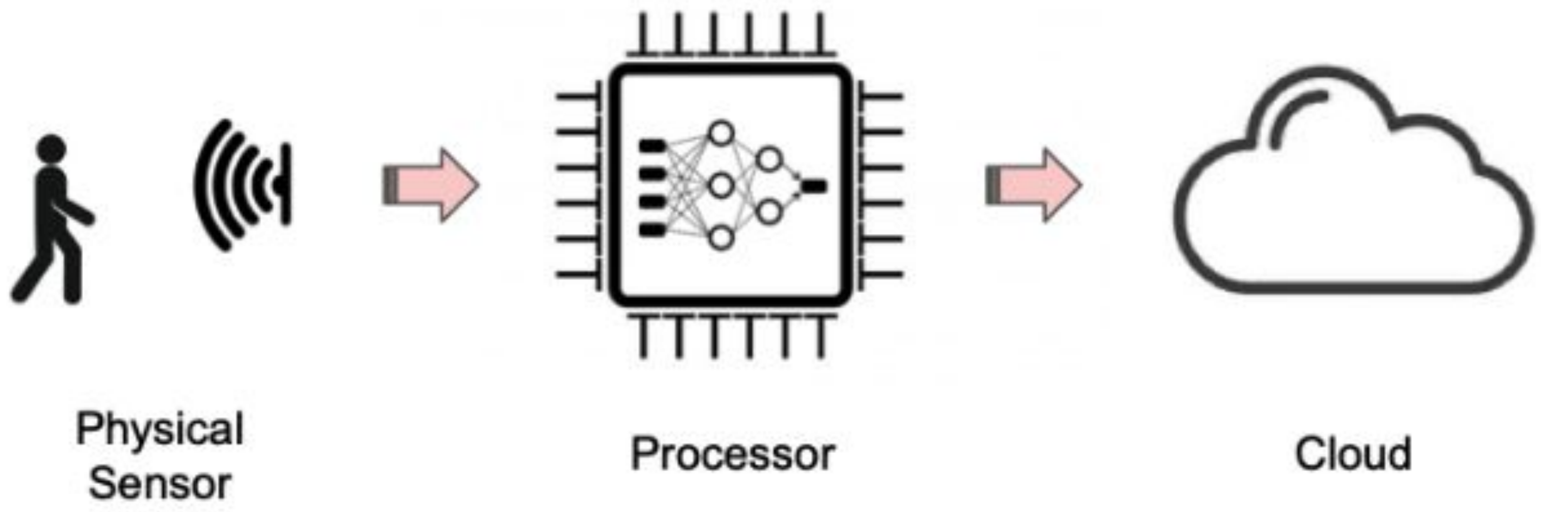
WALC 2024
Applied AI

Applied AI Track Wrap-up

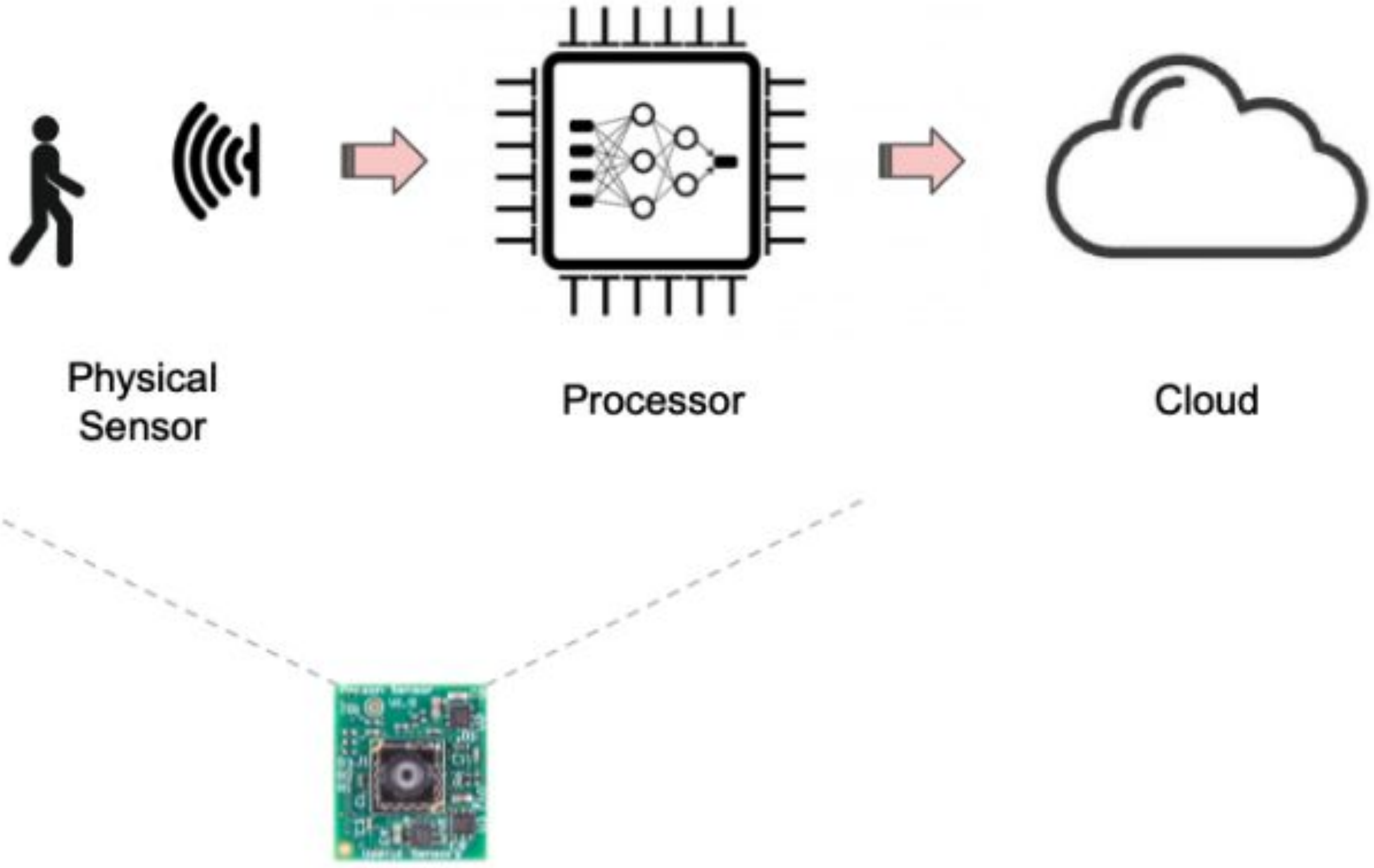
Prof. Marcelo J. Rovai
rovai@unifei.edu.br

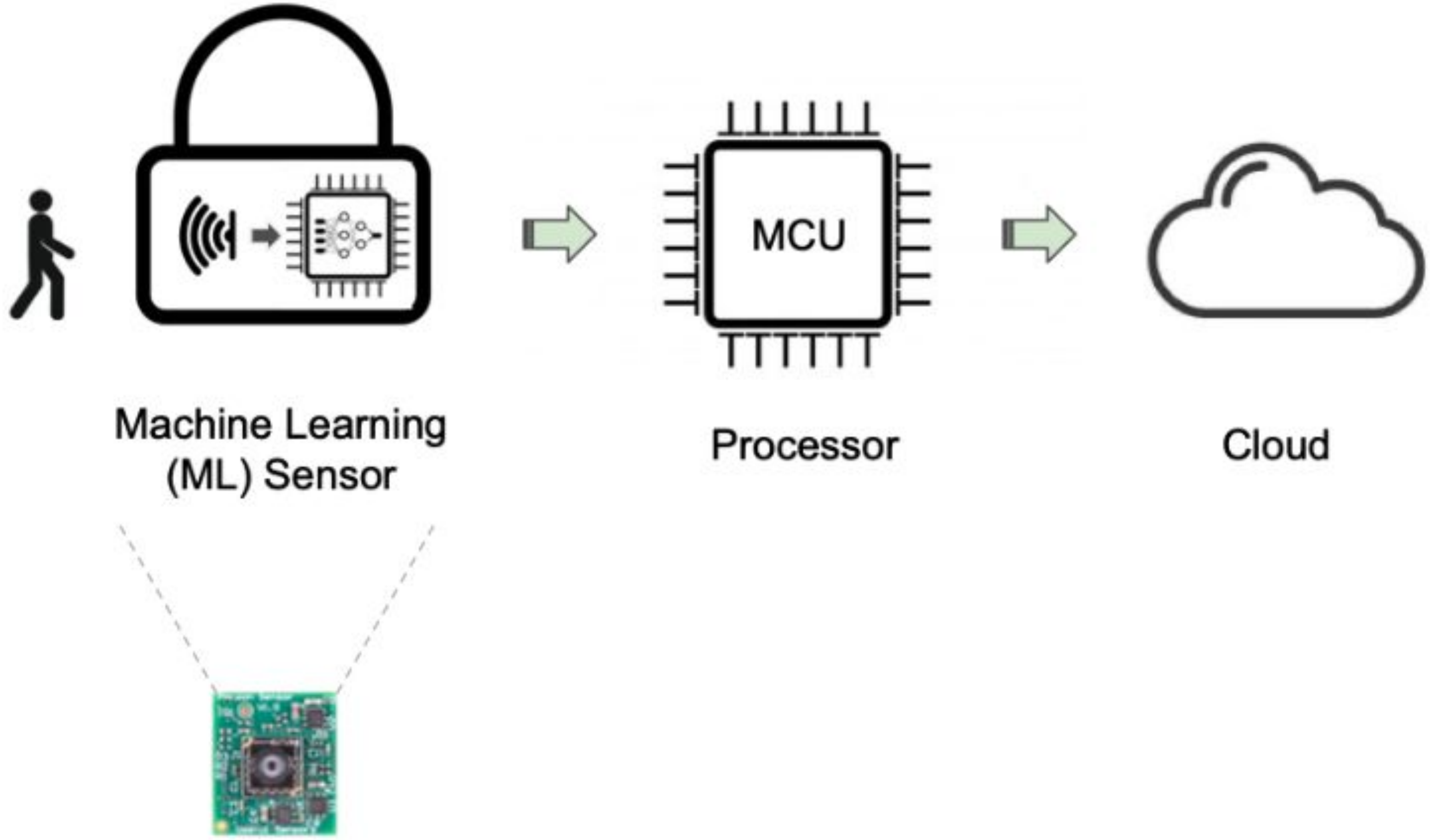
UNIFEI - Federal University of Itajuba, Brazil
TinyML4D Academic Network Co-Chair

The Future of the EdgeAI



Sensor 1.0


Physical Sensor
Processor
Cloud



Sensor 2.0

useful sensors

Person detector

Gaze sensor

Voice command

Text recognizer

...

The Person Sensor has built in facial recognition and determines how many people there are, as well as their relative position.

[USD 10 -> https://www.sparkfun.com/products/21231](https://www.sparkfun.com/products/21231)

mlsensors.org

https://github.com/harvard-edge/ML-Sensors

MACHINE LEARNING SENSORS

Pete Warden [1] Matthew Stewart [2] Brian Plancher [2] Colby Banbury [2] Shvetank Prakash [2] Emma Chen [2]
Zain Asgar [1] Sachin Katti [1] Vijay Janapa Reddi [2]

[1]Stanford University [2]Harvard University

ABSTRACT

Machine learning sensors represent a paradigm shift for the future of embedded machine learning applications. Current instantiations of embedded machine learning (ML) suffer from complex integration, lack of modularity, and privacy and security concerns from data movement. This article proposes a more data-centric paradigm for embedding sensor intelligence on edge devices to combat these challenges. Our vision for "sensor 2.0" entails segregating sensor input data and ML processing from the wider system at the hardware level and providing a thin interface that mimics traditional sensors in functionality. This separation leads to a modular and easy-to-use ML sensor device. We discuss challenges presented by the standard approach of building ML processing into the software stack of the controlling microprocessor on an embedded system and how the modularity of ML sensors alleviates these problems. ML sensors increase privacy and accuracy while making it easier for system builders to integrate ML into their products as a simple component. We provide examples of prospective ML sensors and an illustrative datasheet as a demonstration and hope that this will build a dialogue to progress us towards sensor 2.0.

1 INTRODUCTION

Since the advent of AlexNet [43], deep neural networks have proven to be robust solutions to many challenges that involve making sense of data from the physical world. Machine learning (ML) models can now run on low-cost, low-power hardware capable of deployment as part of an embedded device. Processing data close to the sensor on an embedded device allows for an expansive new variety of always-on ML use-cases that preserve bandwidth, latency, and energy while improving responsiveness and maintaining data privacy. This emerging field, commonly referred to as embedded ML or tiny machine learning (TinyML) [73, 18, 39, 59], is paving the way for a prosperous new array of use-cases, from personalized health initiatives to improving manufacturing productivity and everything in-between.

However, the current practice for combining inference and sensing is cumbersome and raises the barrier of entry to embedded ML. At present, the general design practice is to design or leverage a board with decoupled sensors and compute (in the form of a microcontroller or DSP), and for the developer to figure out how to run ML on these embedded platforms. The developer is expected to train and optimize ML models and fit them within the resource constraints of the embedded device. Once an acceptable prototype implementation is developed, the model is integrated with the rest of the software on the device. Finally, the widget is tethered to the device under test to run inference. The current approach is slow, manual, energy-inefficient, and error-prone.

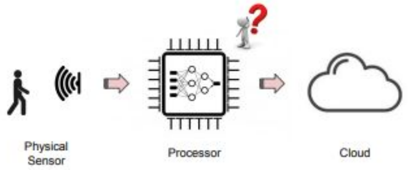


Figure 1. The Sensor 1.0 paradigm tightly couples the ML model with the application processor and logic, making it difficult to provide hard guarantees about the ML sensor's ultimate behavior.

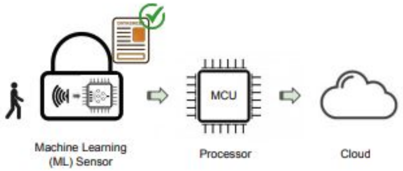
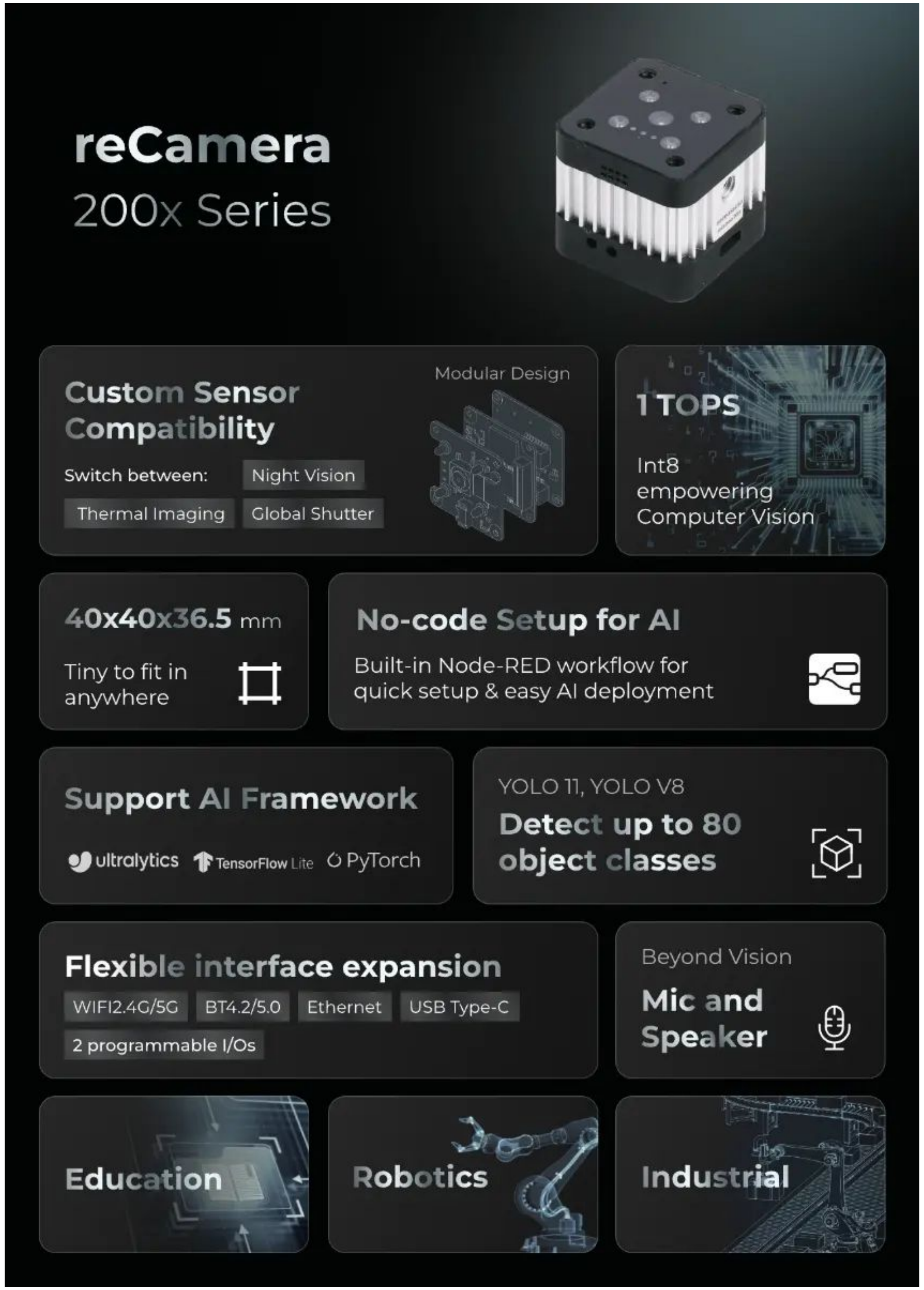


Figure 2. Our proposed Sensor 2.0 paradigm. The ML model is tightly coupled with the physical sensor, separate from the application processor, and comes with an ML sensor datasheet that makes its behavior transparent to the system integrators and developers.

It requires a sophisticated understanding of ML and the intricacies of ML model implementations to optimize and fit a model within the constraints of the embedded device.



seeed studio

https://www.seeedstudio.com/reCamera-2002w-8GB-p-6250.html

Raspberry Pi

12.3 MP Sony IMX500 Intelligent Vision Sensor with a powerful neural network accelerator

Framerates:

- 2×2 binned: 2028×1520 10-bit 30fps
- Full resolution: 4056×3040 10-bit 10fps

7.857 mm sensor size

1.55 μm × 1.55 μm pixel size

78.3 (±3) degree FoV with manual/mechanical adjustable focus

F1.79 focal ratio

25 × 24 × 11.9 mm module dimensions

Integrated RP2040 for neural network firmware management

Works with all Raspberry Pi models, using our standard camera connector cable

https://www.raspberrypi.com/documentation/accessories/ai-camera.html

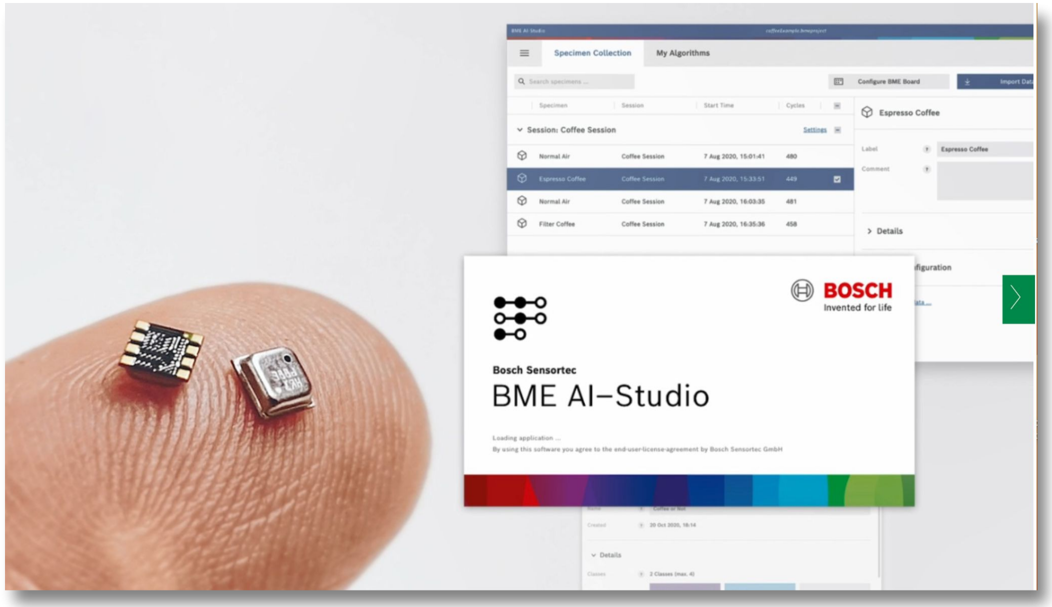
Bosch BME688 - Environmental sensing with AI

Relative humidity barometric pressure

Excellent temperature stability

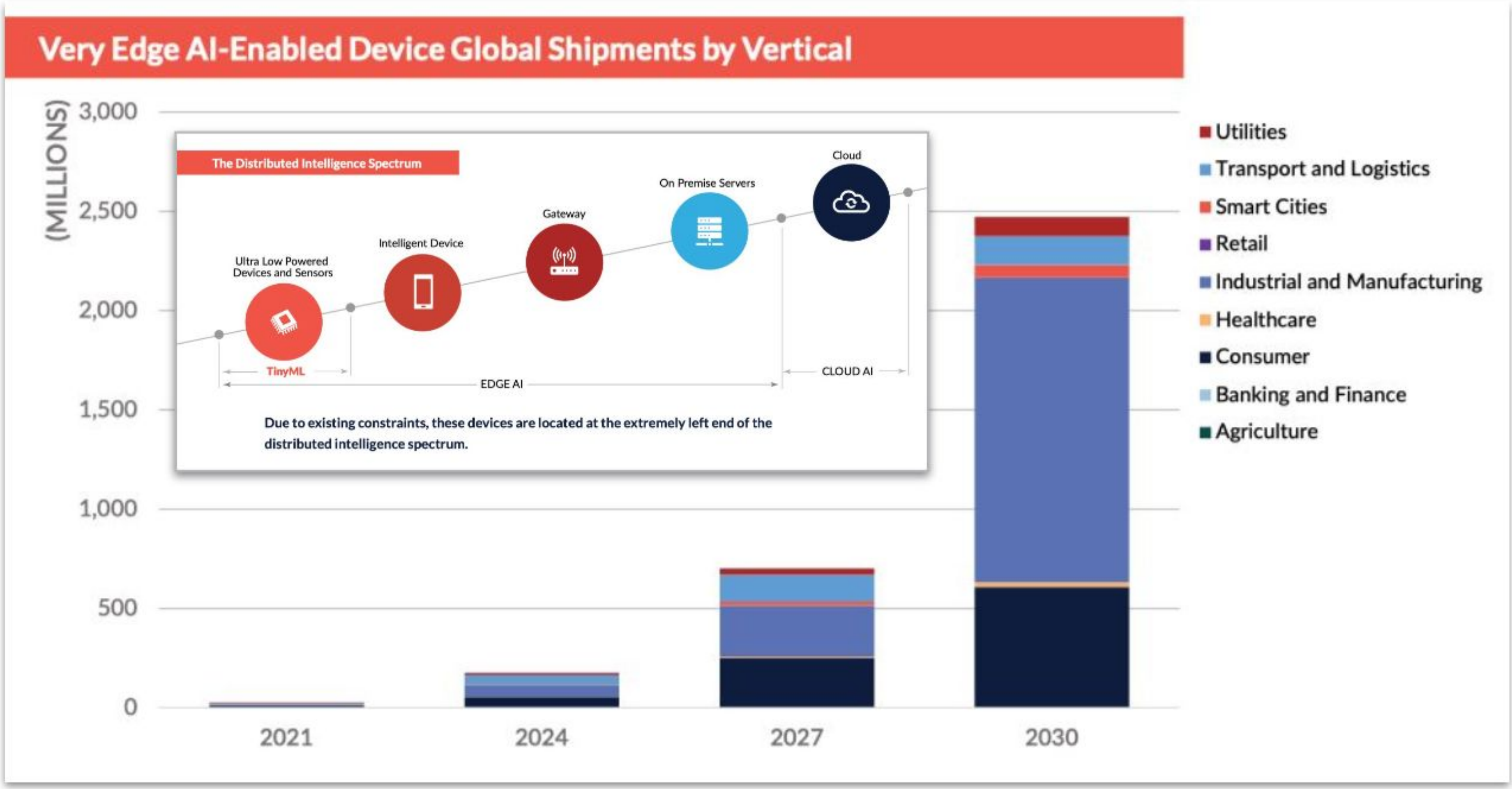
Humidity

Gas sensing



https://www.bosch-sensortec.com/products/environmental-sensors/gas-sensors/bme688/

Massive Potential for Impact



Source: ABI Research: TinyML

microsoft/**BitNet**

Official inference framework for 1-bit LLMs

Bitnet.cpp employs one-bit quantization, representing values with a ternary system (+1, -1, 0). This approach simplifies calculations by replacing complex multiplications with additions and subtractions, eliminating the need for GPUs.

- Speedups range from 1.37x to 6.1x on various CPUs.
- Power consumption reductions between 55.4% and 82.2% compared to traditional GPU-based inference.

bitnet.cpp

LLAMA 3.2 1B on Arm CPU with Meta's Executorch

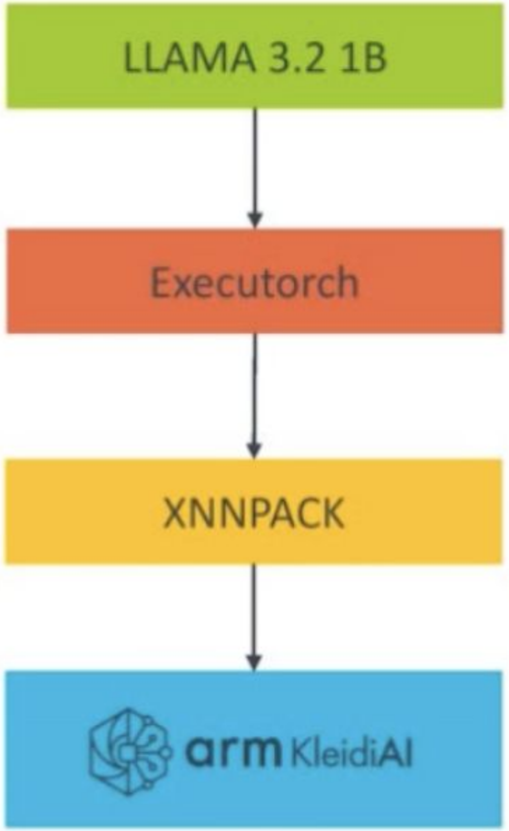
Best in class LLM performance on Arm

LLM Chatbot on Vivo X100
(4 x CPU Threads)

GEMMA 2B **Tok/s** (higher is better)		
	2 x Threads	4 x Threads
Prompt / TTFT phase	218	350
Text Generation phase	42	**50**

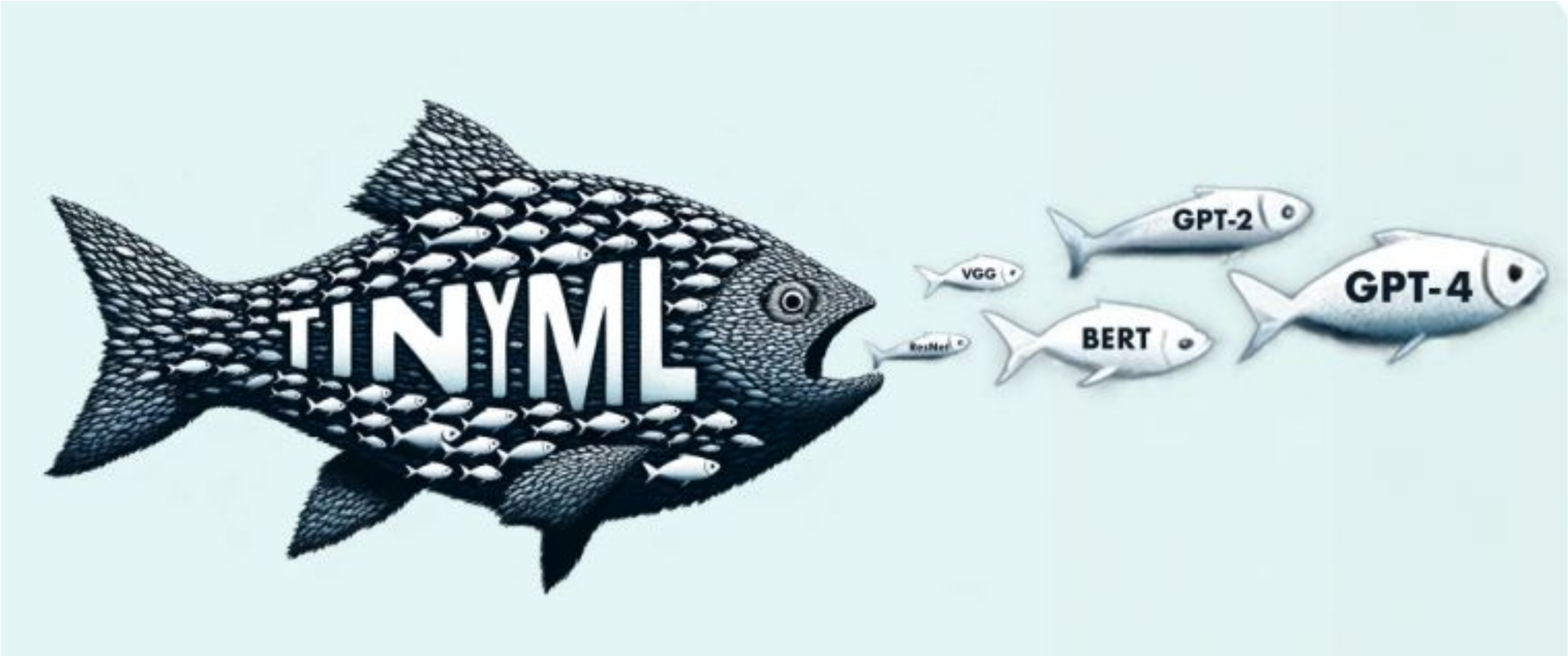
25% - 30% Uplift to LLAMA 3.2 1B
when using KleidiAI

arm KleidiAI



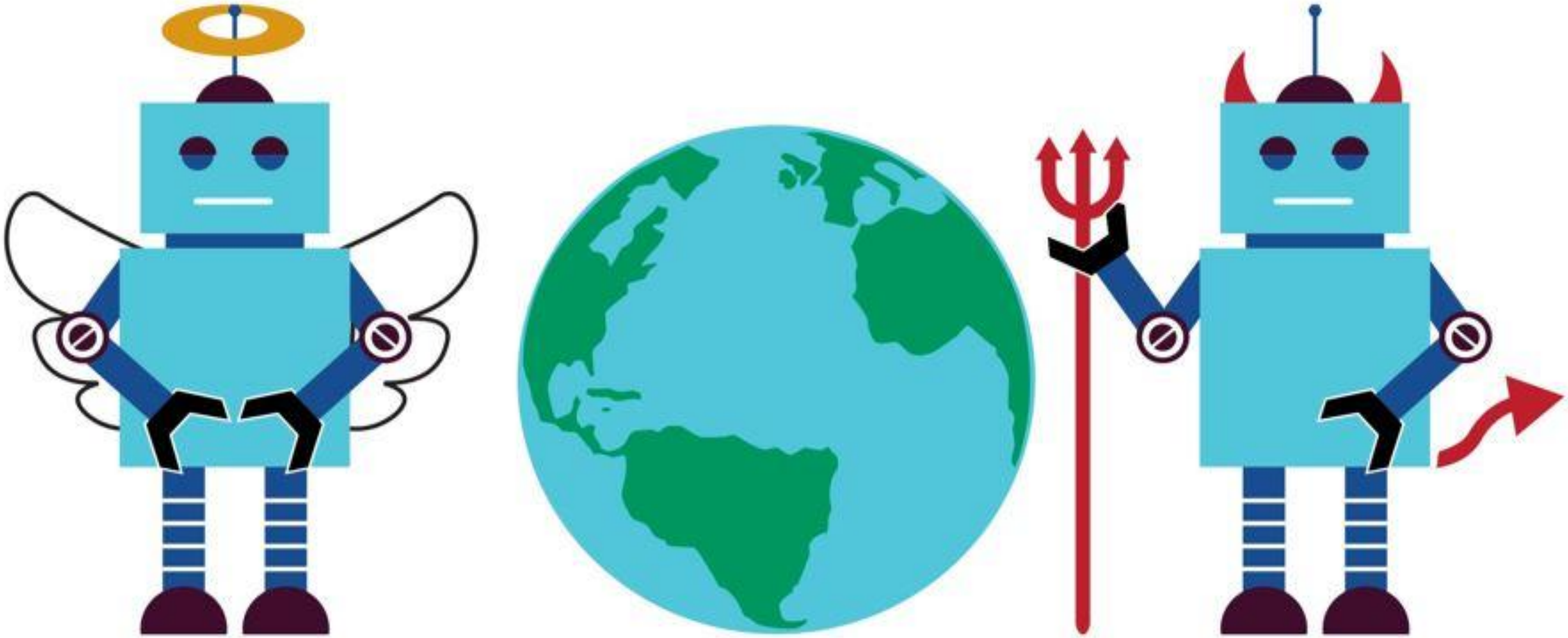
Edge AI has a very high impact potential, and it is for now!

TinyML: Why the Future of Machine Learning is Tiny and Bright



Shvetank Prakash, Emil Njor, Colby Banbury, Matthew Stewart, Vijay Janapa Reddi

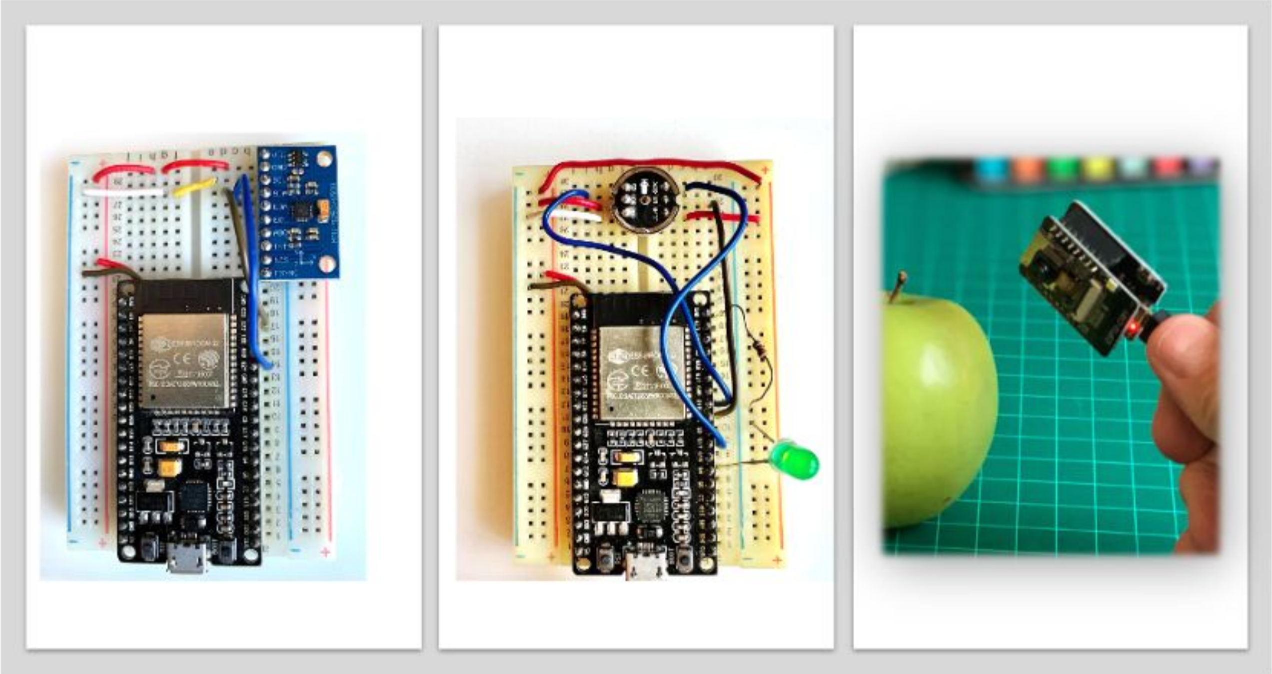
Responsible AI



To learn more ...

ESP32-TinyML

Exploring TinyML with ESP32 MCUs.



Seeed-XIAO-BLE-Sense

KWS, Anomaly Detection & Motion Classification and Micropython - Exploring the Seeed XIAO BLE Sense.

Programming Tiny devices with MicroPython. The easiest way!
MJRoBot (Marcelo Rovai)

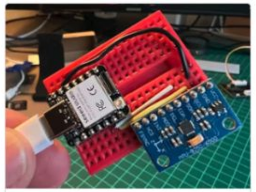
Sensor DataLogger
MJRoBot (Marcelo Rovai)

TinyML Made Easy: Anomaly Detection & Motion Classification
MJRoBot (Marcelo Rovai)

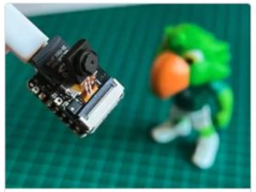
TinyML Made Easy: Sound Classification (KWS)
MJRoBot (Marcelo Rovai)

XIAO-ESP32S3-Sense

TinyML Made Easy: KeyWord Spotting (KWS)
MJRoBot (Marcelo Rovai)



Exploring Machine Learning with the new XIAO ESP32S3
MJRoBot (Marcelo Rovai)



TinyML Made Easy: Image Classification
MJRoBot (Marcelo Rovai)

To learn more …

Online Courses

[Harvard School of Engineering and Applied Sciences - CS249r: Tiny Machine Learning]
Professional Certificate in Tiny Machine Learning (TinyML) – edX/Harvard
Introduction to Embedded Machine Learning - Coursera/Edge Impulse
Computer Vision with Embedded Machine Learning - Coursera/Edge Impulse
UNIFEI-IESTI01 TinyML: "Machine Learning for Embedding Devices"

Books

"Python for Data Analysis" by Wes McKinney
"Deep Learning with Python" by François Chollet - GitHub Notebooks
"TinyML" by Pete Warden and Daniel Situnayake
"TinyML Cookbook 2nd Edition" by Gian Marco Iodice
"Technical Strategy for AI Engineers, In the Era of Deep Learning" by Andrew Ng
"AI at the Edge" book by Daniel Situnayake and Jenny Plunkett
"XIAO: Big Power, Small Board" by Lei Feng and Marcelo Rovai
"MACHINE LEARNING SYSTEMS for TinyML" by a collaborative effort

Projects Repository

Edge Impulse Expert Network

On the TinyML4D website, You can find lots of educational materials on TinyML. They are all free and open-source for educational uses – we ask that if you use the material, please cite them! TinyML4D is an initiative to make TinyML education available to everyone globally.

TinyML4D **Show&Tell** Presentations

[TinymML4D Academic Network Show and Tell Main Index.]

The TinyML4D Academic Network Students should use this form to propose presentations.
https://forms.gle/ic52HZMqVv4pBrkP7 2

The Show and Tell are typically held at 2 pm UTC on the last Thursday of each month and will take place in this Meet link:
https://meet.google.com/rns-yyrx-ggw

TINYML4D

Projects by Students (UNIFEI – IESTI01)

- **Sound:**
  - Earthquake detection
  - Covid Detection (cough)
  - Key Detection
  - Pulmonary Disease
  - Snore Detection
  - Bionic Hand Control

- **Other Sensors:**
  - Bionic Hand – Finger Detection
  - Electric Charges
  - ECG – Fibrial Atrilation detection

- **Image:**
  - Mask Detection
  - Forest Fire Detection
  - Helmet Detection
  - Water Consumption (hydrometer)
  - Sign Language
  - Coffee Disease Classification
  - Bee Counting

- **Vibration:**
  - Personal Trainer
  - Bearing – Anomaly Detection

UNIFEI IESTI01 2021.1 – Final Projects   UNIFEI IESTI01 2021.2 – Final Projects   UNIFEI IESTI01 2022.1 – Final Projects

Questions?

TINYML4D