

Tutorial: On-device AI

CS4262/5462: Machine Learning Systems

Junyi Shen

junyis@u.nus.edu

17 Feb 2026

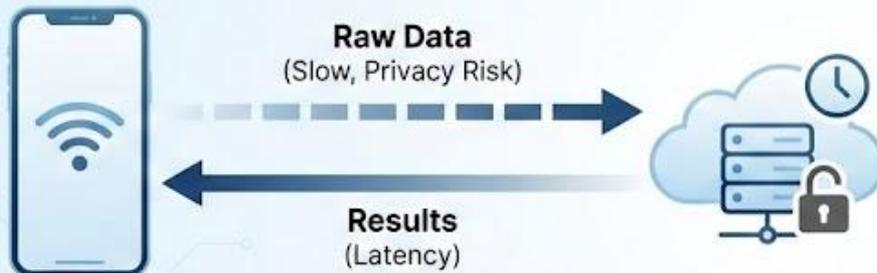
01

On-device AI

On-device AI: Intelligent Experiences, Redefined

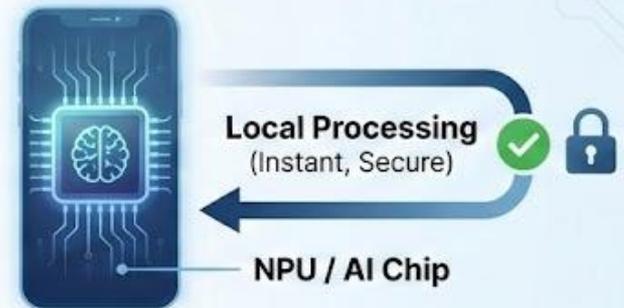
Processing data directly on local hardware for instant, secure capability

CLOUD AI
(Traditional)



Dependent on connectivity, data leaves device

ON-DEVICE AI
(Modern)



Runs offline, data stays local



Smartphones:
Face ID, Photo Assist



Wearables:
Health Monitoring



Automotive:
Driver Assist

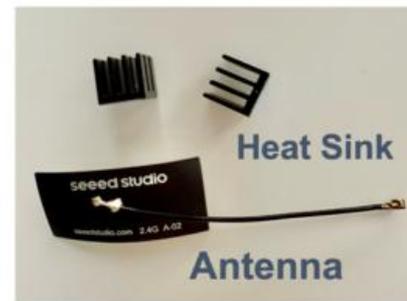
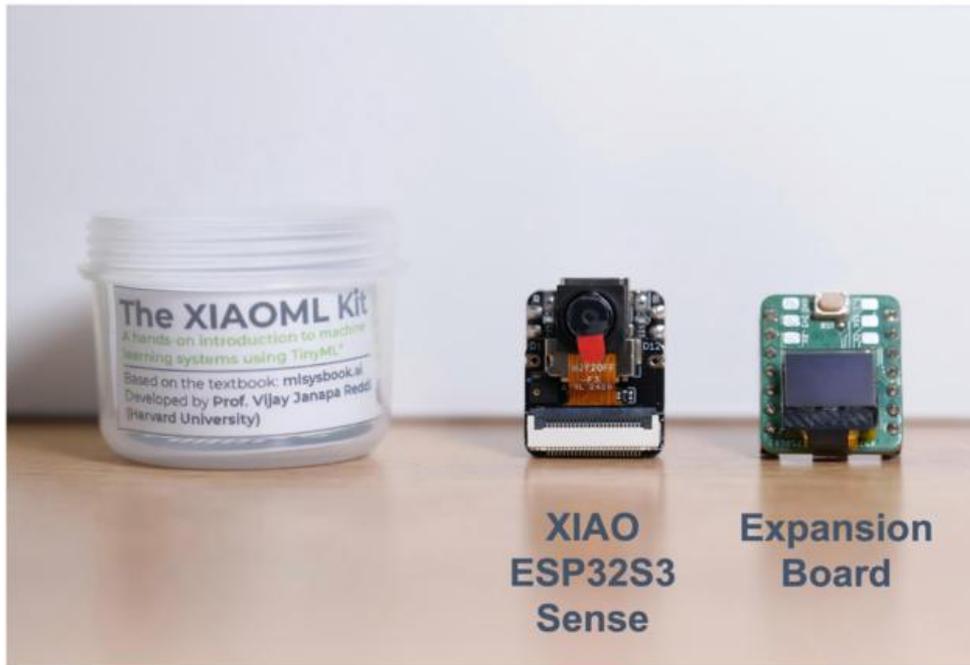


Smart Home:
Local Voice Commands

01

Device

- **XIAO ESP32S3 Sense:** Main development board with integrated camera sensor, digital microphone, and SD card support
- **Expansion Board:** Features a 6-axis IMU (LSM6DS3TR-C) and 0.42" OLED display for motion sensing and data visualization
- **SD Card Toolkit:** Includes SD card and USB adapter for data storage and model deployment
- **USB-C Cable:** For connecting the board to your computer
- **Antenna and Heat Sinks**



01

Install

We use Arduino for edge-AI programming.

The screenshot displays the Arduino IDE 2.3.7 interface. The top bar shows the title "sketch_jan19a | Arduino IDE 2.3.7". The left sidebar contains the "BOARDS MANAGER" with a search filter set to "ESP32". Under "Arduino ESP32 Boards by Arduino", the version "2.0.18" is selected, and an "INSTALL" button is visible. Below that, "esp32 by Espressif Systems" is shown with version "2.0.17" installed, and a "REMOVE" button is present. The main editor area shows a sketch named "sketch_jan19a.ino" with the following code:

```
1 void setup() {
2   // put your setup code here, to run once:
3
4 }
5
6 void loop() {
7   // put your main code here, to run repeatedly:
8
9 }
10
```

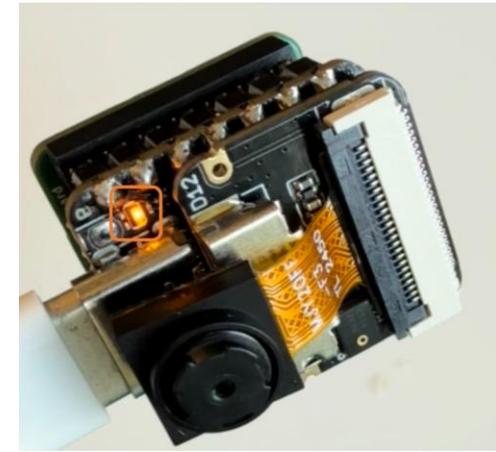
The "Output" window at the bottom shows the following log:

```
esp32: riscv32-esp-elf-gdb@11.2_20220823 installed
Installing esp32:xtensa-esp-elf-gdb@11.2_20220823
Configuring tool.
esp32:xtensa-esp-elf-gdb@11.2_20220823 installed
Installing esp32:xtensa-esp32-elf-gcc@esp-2021r2-patch5-8.4.0
Configuring tool.
esp32:xtensa-esp32-elf-gcc@esp-2021r2-patch5-8.4.0 installed
Installing esp32:xtensa-esp32s2-elf-gcc@esp-2021r2-patch5-8.4.0
Configuring tool.
esp32:xtensa-esp32s2-elf-gcc@esp-2021r2-patch5-8.4.0 installed
Installing esp32:xtensa-esp32s3-elf-gcc@esp-2021r2-patch5-8.4.0
Configuring tool.
esp32:xtensa-esp32s3-elf-gcc@esp-2021r2-
Installing platform esp32:esp32@2.0.17
Configuring platform.
Platform esp32:esp32@2.0.17 installed
```

A notification box at the bottom right states: "Successfully installed platform esp32:2.0.17". The status bar at the bottom indicates "Offline" and "Ln 1, Col 1 x No board selected".

01

BLINK



```
#define LED_BUILT_IN 21 // This line is optional
```

```
void setup() {  
  pinMode(LED_BUILT_IN, OUTPUT); // Set the pin as output  
}
```

Run at Start

```
// Remember that the pins work with inverted logic  
// LOW to turn on and HIGH to turn off
```

```
void loop() {  
  digitalWrite(LED_BUILT_IN, LOW); //Turn on  
  delay (1000); //Wait 1 sec  
  digitalWrite(LED_BUILT_IN, HIGH); //Turn off  
  delay (1000); //Wait 1 sec  
}
```

Run in Loop

Note that the pins operate with inverted logic: LOW turns on and HIGH turns off.

01 Microphone

XIAOML Toolkit equipped with a digital microphone.

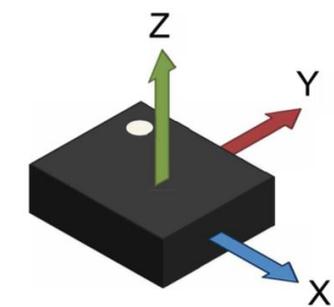


https://github.com/Mjrovai/XIAO-ESP32S3-Sense/tree/main/Mic_Test/XiaoEsp32s3_Mic_Test

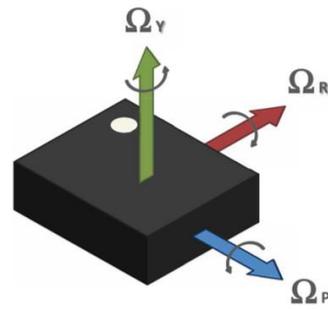
01 IMU Sensor

An Inertial Measurement Unit (IMU) is a sensor that measures motion and orientation. The LSM6DS3TR-C on your XIAOML kit is a 6-axis IMU, meaning it combines:

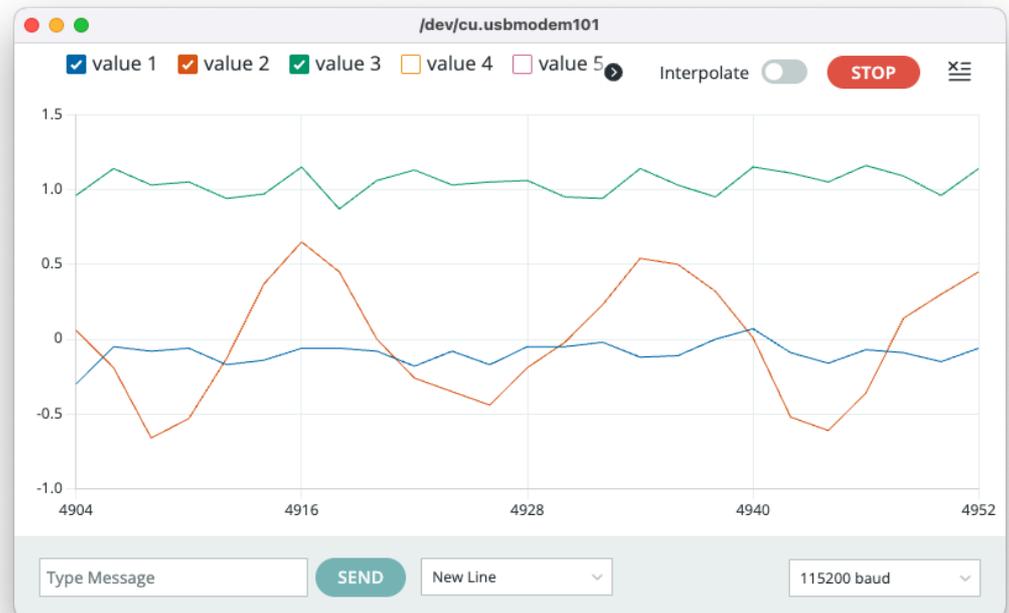
- **3-axis Accelerometer:** Measures linear acceleration (including gravity) along X, Y, and Z axes
- **3-axis Gyroscope:** Measures angular velocity (rotation rate) around X, Y, and Z axes



Direction of detectable acceleration (top view)



Direction of detectable angular rate (top view)



01 WIFI

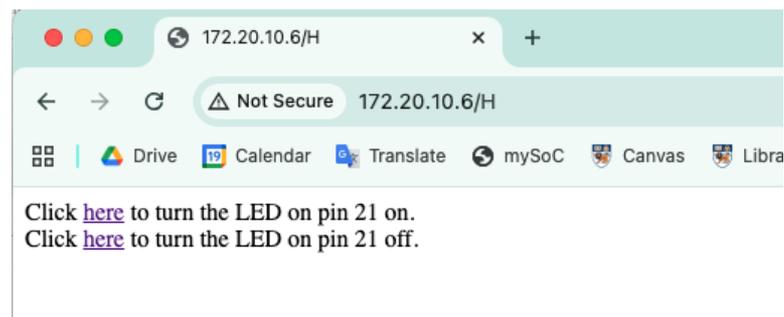
WIFI Scan and remote control

```
Scan start
Scan done
36 networks found
Nr | SSID | RSSI | CH | Encryption
1 | Junyi-iPhone | -30 | 3 | WPA2
2 | eduroam | -56 | 3 | WPA2-EAP
3 | NUS | -57 | 3 | WPA2-EAP
4 | NUS_STU | -57 | 3 | WPA2-EAP
5 | SOCLAB | -57 | 3 | WPA+WPA2
6 | NUS_Guest | -57 | 3 | open
7 | eduroam | -59 | 9 | WPA2-EAP
8 | NUS_STU | -59 | 9 | WPA2-EAP
9 | SOCLAB | -59 | 9 | WPA+WPA2
```

```
Connecting to Junyi-iPhone
..
WiFi connected.
IP address:
172.20.10.6
```

7:15   4G 99

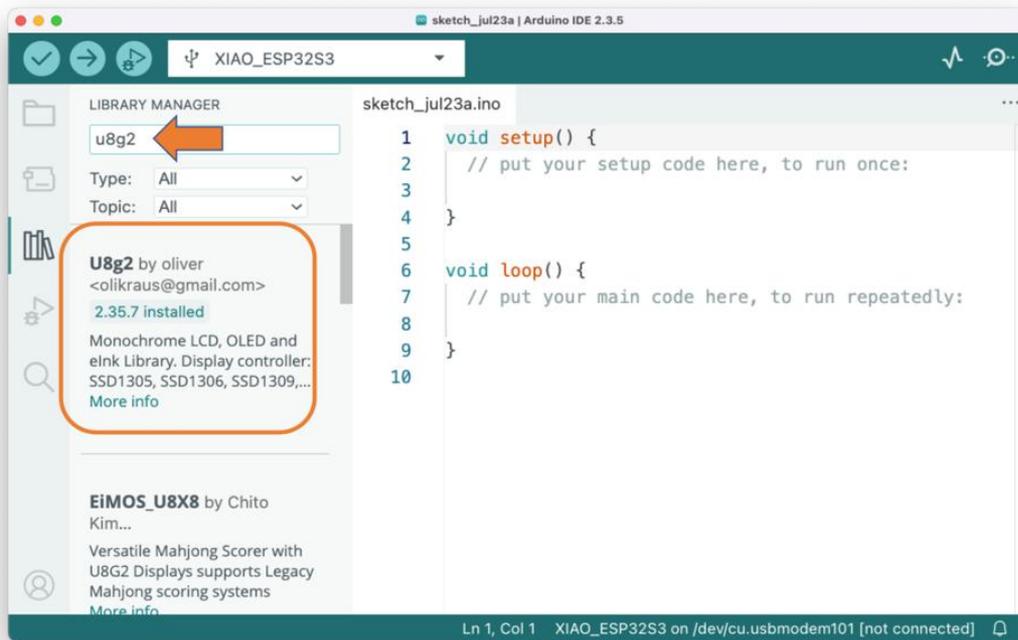
Click [here](#) to turn the LED on pin 21 on.
Click [here](#) to turn the LED on pin 21 off.



01

OLED

A good way to demonstrate your output – OLED



Test Code:

https://mlsysbook.ai/kits/contents/seeed/xiao_esp32s3/setup/setup.html#sec-setup-test-code-998c

02 Project Topics

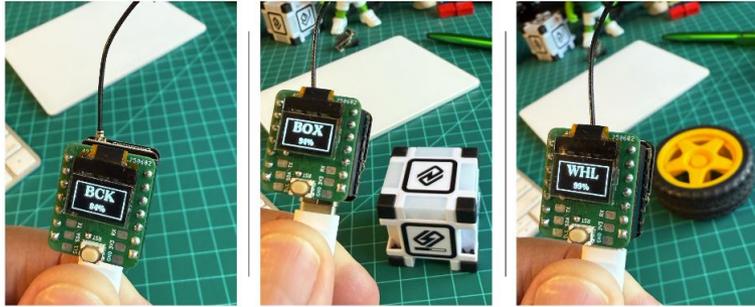


Image Classification

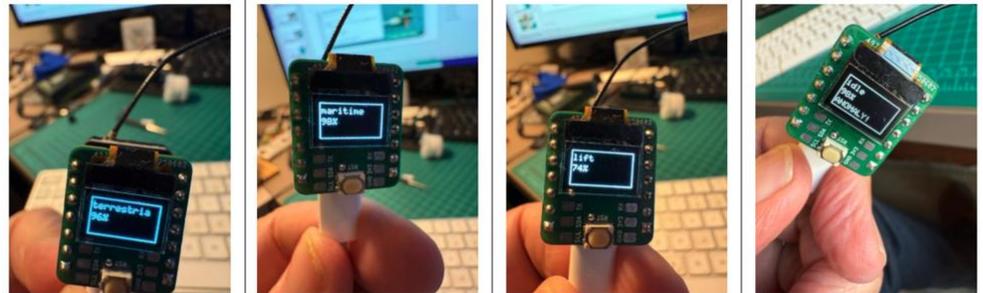


Object Detection



Keyword Spotting

**Motion Classification
Anomaly Detection**



02

Project Requirements

Modality	Task	Description
Vision	Image Classification	Learn to classify images
Vision	Object Detection	Implement object detection
Sound	Keyword Spotting	Explore voice recognition systems
IMU	Motion Classification and Anomaly Detection	Classify motion data and detect anomalies

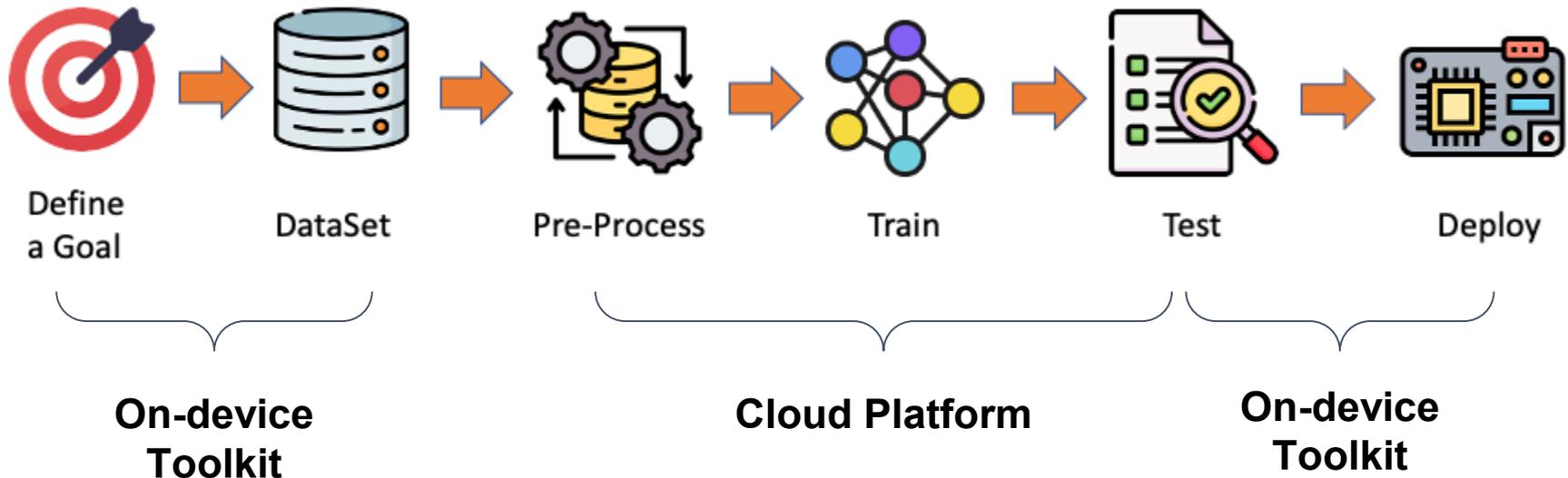
Requirements:

**2 of the tutorial tasks from the above (different setting);
1 in-depth exploration with advanced post-processing**

Examples:

https://mlsysbook.ai/kits/contents/seeed/xiao_esp32s3/xiao_esp32s3.html

02 Project Workflows



Arduino: On-device programming

SenseCraft AI: Easy to use, but less flexibility

Edge Impulse Studio: More powerful, seamless with Arduino

02 Project Workflows



Define
a Goal

Imagine you are creators of a startup of On-device AI...

Which setting should we focus on?

Industrial manufacturing, elderly/child caring...

What functions should we design?

Image classification, voice recognition...

How can we implement these functions leveraging the device we have?

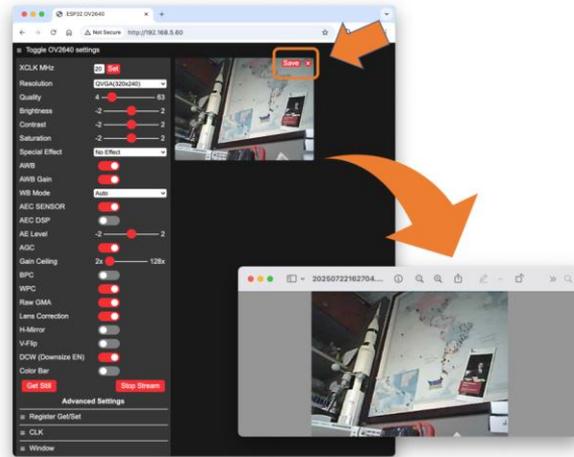
Camera, IMU unit, Microphone, WiFi, LED...

02 Project Workflows

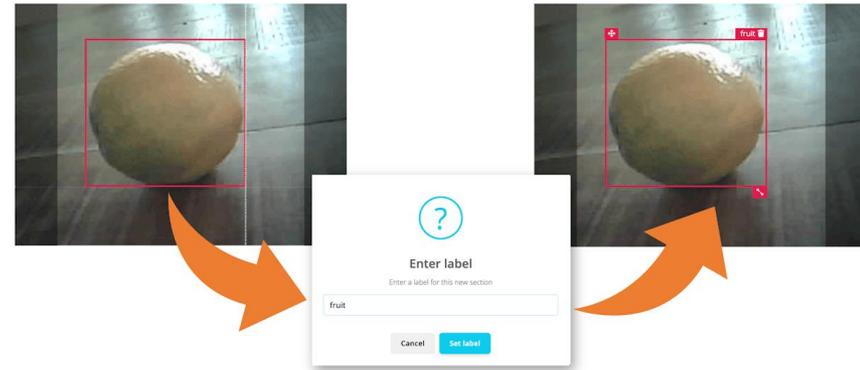


DataSet

Gather data for your model training.



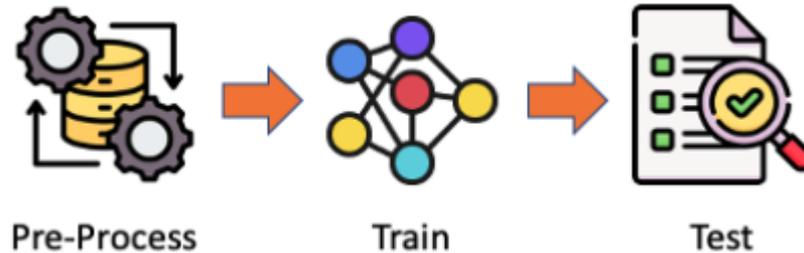
Gathering (Device/Phone)



Labeling and Uploading

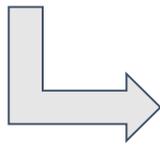
https://mlsysbook.ai/kits/contents/seeed/xiao_esp32s3/object_detection/object_detection.html#sec-object-detection-collecting-dataset-xiao-esp32s3-9814

02 Project Workflows



**Model Choose
Hyper-parameter Config**

Test your trained model

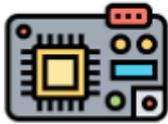


Training (On platform)



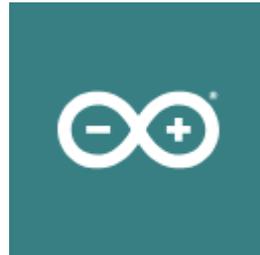
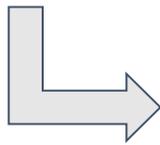
https://mlsysbook.ai/kits/contents/seeed/xiao_esp32s3/image_classification/image_classification.html#sec-image-classification-preprocessing-feature-generation-ee2e

02 Project Workflows

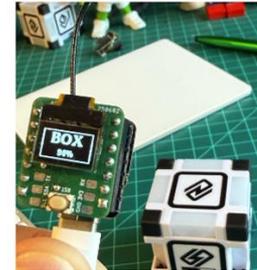


Deploy

**Download the
trained model**



**Migrate to Edge device
via Arduino**



**Test with your
edge device**



https://mlsysbook.ai/kits/contents/seeed/xiao_esp32s3/image_classification/image_classification.html#sec-image-classification-model-deployment-arduino-library-ei-studio-5579

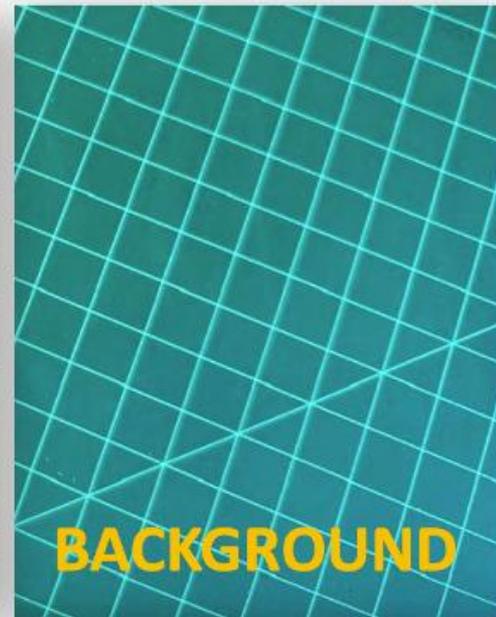
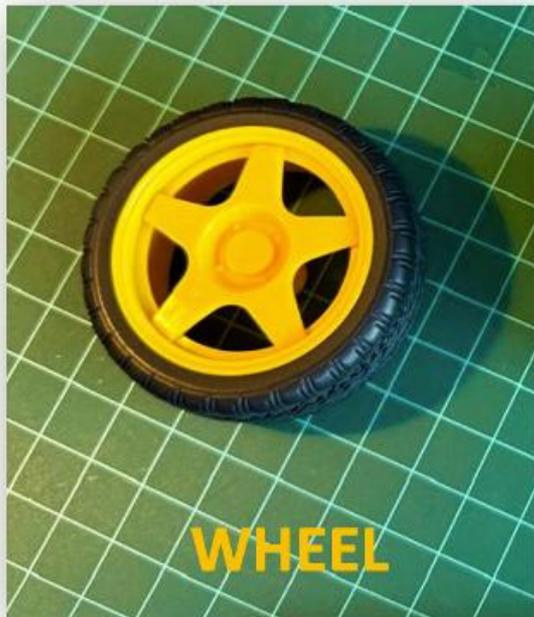
02 Project Example

An industrial installation that should automatically sort wheels and boxes.



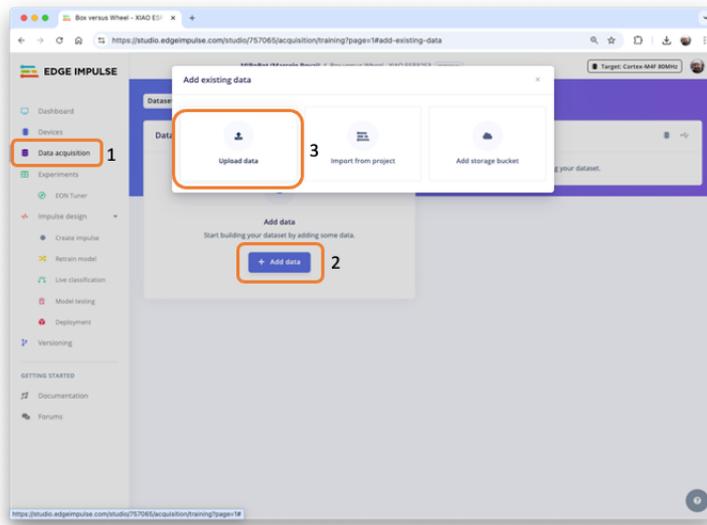
02 Project Example

An industrial installation that should automatically sort wheels and boxes.

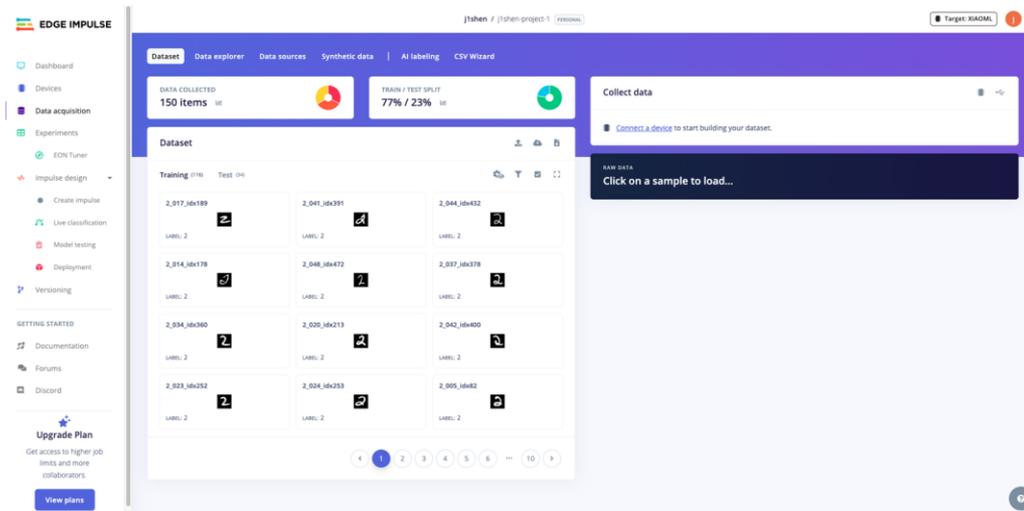


02 Project Example

Upload your dataset to the EDGE IMPULSE Platform

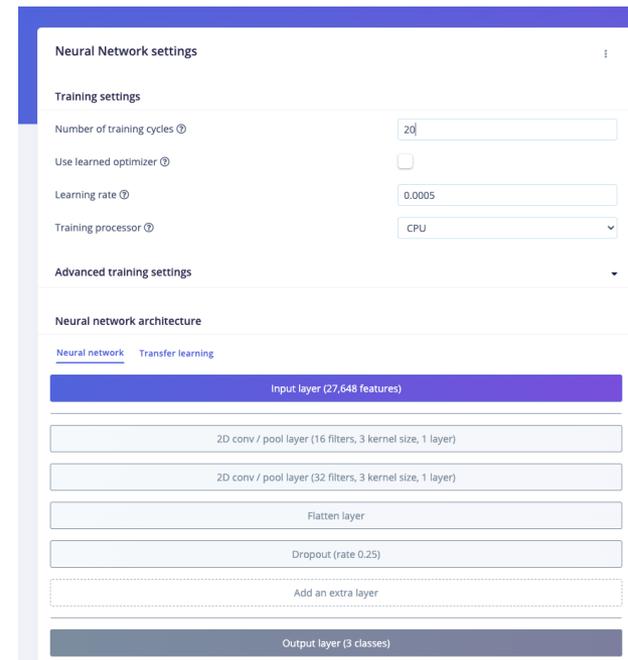
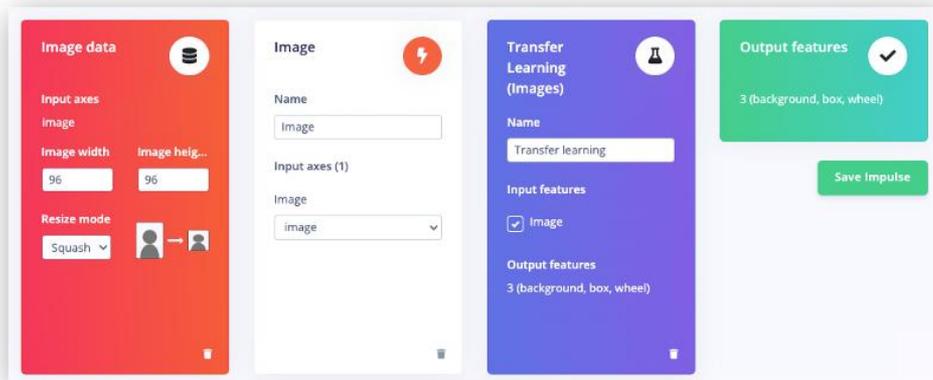


Labeling



02 Project Example

Define training model and parameters



02 Project Example

Train/Test the model and get the report.

The image shows a two-panel interface for training and testing a neural network model. The left panel displays configuration settings, and the right panel shows the resulting performance report.

Neural Network settings

Training settings

- Number of training cycles: 20
- Use learned optimizer:
- Learning rate: 0.0005
- Training processor: CPU
- Data augmentation:

Advanced training settings

- Validation set size: 20
- Split train-validation set on metadata key:
- Batch size: 32
- Auto-weight classes:
- Profile int8 model:

Neural network architecture

- Input layer (27,648 features)
- MobileNetV2_96x96_0.35 (final layer: 16 neurons, 0.1 dropout)
- Choose a different model
- Output layer (3 classes)

Model (Model version: Quantized QINT8)

Last training performance (validation set)

- ACCURACY: 100.0%
- LOSS: 0.01

Confusion matrix (validation set)

	BACKGROUND	DISK	WHEEL
BACKGROUND	100%	0%	0%
DISK	0%	100%	0%
WHEEL	0%	0%	100%
F1 SCORE	1.00	1.00	1.00

Metrics (validation set)

Metric	Value
Area under ROC Curve	1.00
Weighted average Precision	1.00
Weighted average Recall	1.00
Weighted average F1 score	1.00

Data explorer (full training set)

- background - correct
- disk - correct
- wheel - correct

On-device performance (Engine: Qualcomm Compiler (ARM optimized))

- INFERRING TIME: 1160 ms.
- PEAK RAM USAGE: 232.9K
- FLASH USAGE: 546.2K

02 Project Example

Deploy the model to edge device and implement the post-processing functions.

The screenshot shows the 'Configure your deployment' page in Edge Impulse Studio. It features a search bar for the Arduino library, a 'SELECTED DEPLOYMENT' section for the Arduino library, and 'MODEL OPTIMIZATIONS' for TensorFlow Lite. Two tables compare quantized and unquantized models. An orange arrow points to the 'Run model testing' button.

Quantized (INT8)		IMAGE	TRANSFER LEARNING	TOTAL
Latency	7ms		1,393ms	1,400ms
RAM	40K		353.7K	353.7K
FLASH	-		684.2K	-
ACCURACY	-		-	-

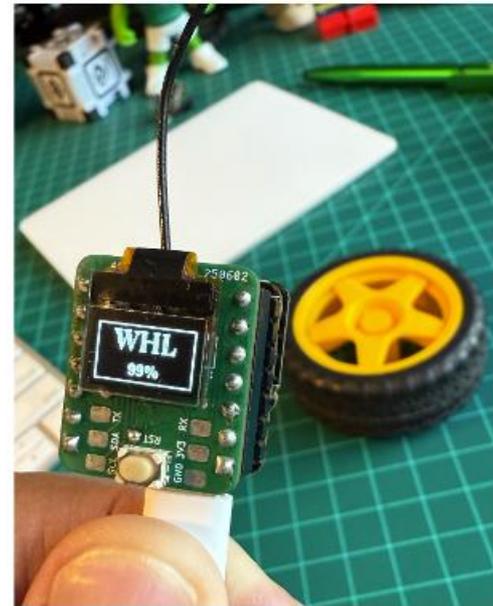
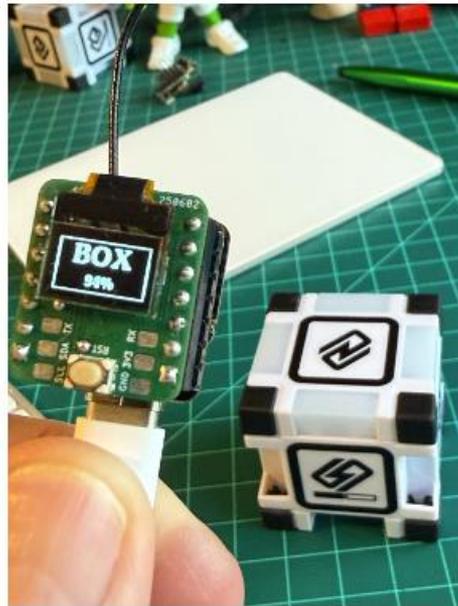
Unquantized (float32)		IMAGE	TRANSFER LEARNING	TOTAL
Latency	7ms		1,557ms	1,564ms
RAM	40K		1.1M	1.1M
FLASH	-		1.7M	-
ACCURACY	-		-	-

The screenshot shows the Arduino IDE Boards Manager for the XIAO_ESP32S3 board. The 'esp32' board is selected, and the 'esp32 by Espressif Systems' package is highlighted. The code editor shows the 'image_class_XIAOAML-Kit.ino' file with preprocessor directives for the board and the TensorFlow Lite model.

```
15 * IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHA
16 * FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EV
17 * AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES O
18 * LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE
19 * OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER
20 * SOFTWARE.
21 */
22
23 // Tested with ESP32 by Espressif Systems Core 2.0.17
24 // on XIAO ESP32S3 Sense V1.0 and v1.1
25 // Adapted from Edge Impulse ESP32 Camera example
26 // Marcelo Rovai, August, 5th 2025
27
28 /* Includes
29 #include <Box_versus_Wheel_-_XIAO_ESP32S3_inferencing.h>
30 #include "edge-impulse-sdk/dsp/image/image.hpp"
31 #include "esp_camera.h"
32
```

02 Project Example

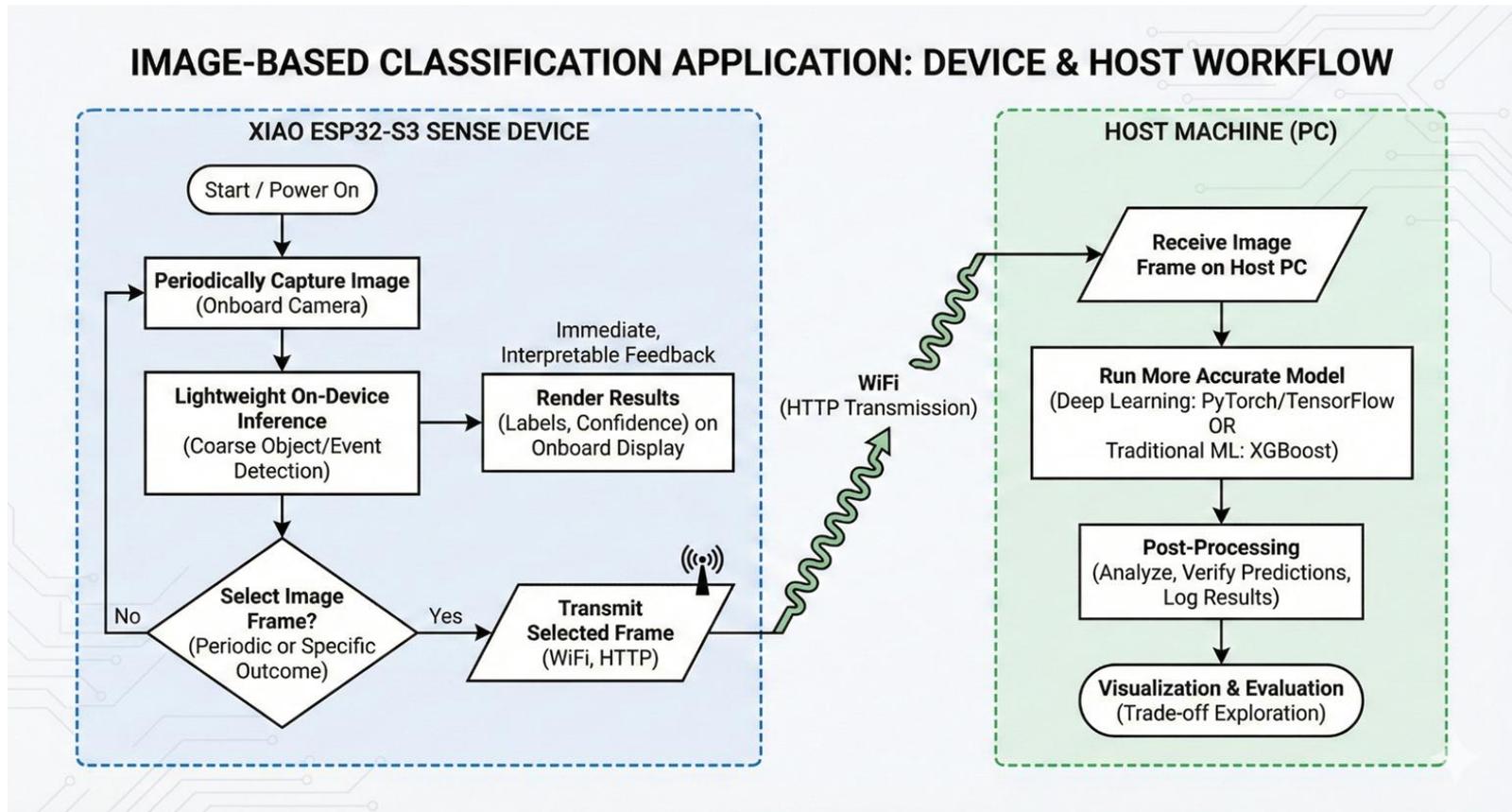
Test the performance of your model.



https://github.com/Mjrovai/XIAO-ESP32S3-Sense/blob/main/XIAOML_Kit_code/XIAOML-Kit-Img_Class_OLED_Gen/XIAOML-Kit-Img_Class_OLED_Gen.ino

02 Project Example

Advanced post-processing and IoT



02 Project Example

ESP32S3 Live Camera



Transmit captured image via WiFi

```
main.py > ...
4 from io import BytesIO
5
6 import torch
7 from transformers import AutoImageProcessor, AutoModelForImageClassification
8
9 # ===== 配置 =====
10 ESP32_URL = "http://192.168.5.214/jpg" # ← 改成你的 ESP32 IP
11 DEVICE = "cuda" if torch.cuda.is_available() else "cpu"
12
13 # ===== 加载 HF 模型 =====
```

```
PS C:\Users\jyshe\OneDrive\桌面\ondevice> uv run .\main.py
pencil sharpener      0.015
mousetrap             0.015

Top-5 Prediction:
mouse, computer mouse 0.864
cassette player       0.024
CD player             0.007
crash helmet          0.006
projector             0.005

Top-5 Prediction:
mouse, computer mouse 0.910
cassette player       0.009
loudspeaker, speaker, speaker unit, loudspeaker system, speaker system 0.005
projector             0.004
crash helmet          0.003

Top-5 Prediction:
mouse, computer mouse 0.851
cassette player       0.022
space heater          0.009
loudspeaker, speaker, speaker unit, loudspeaker system, speaker system 0.009
projector             0.006

Top-5 Prediction:
mouse, computer mouse 0.582
cassette player       0.043
```

Deploy a stronger model on PC

Project Bonus

- **Multi-modality Support:** Design a system that fuses information from more than one sensor simultaneously (e.g., combining Vision and Audio, or Vision and IMU).
- **Hardware Expansion:** Add and integrate new components (sensors, actuators, or servos) to the existing toolkit to enable novel functions.
- **Original Design:** Demonstrate a unique application scenario or system design that significantly differs from the standard course examples.

02 Important Dates

- [Mar 05] Project proposal
- [Mar 26] Mid-term project report
- [Apr 16] Demo presentation
- [Apr 16] Final report

Good luck with your on-device AI journey!



THANK YOU

01

References

<https://mlsysbook.ai/book/>

<https://mlsysbook.ai/kits/>

<https://sensecraft.seeed.cc/ai/home>

<https://www.edgeimpulse.com/>

https://github.com/Mjrovai/XIAO-ESP32S3-Sense/tree/main/XIAOML_Kit_code