life.augmented

STM32 Artificial Intelligence Solutions

MDG Korea

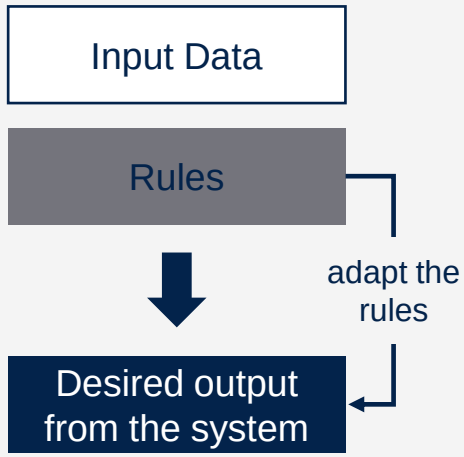
문현수 과장

A new way to add environment awareness to your products

From rule-based engineering to data-driven engineering

Standard programming

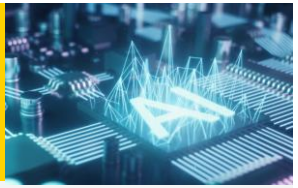
Handcrafted rules based on experience

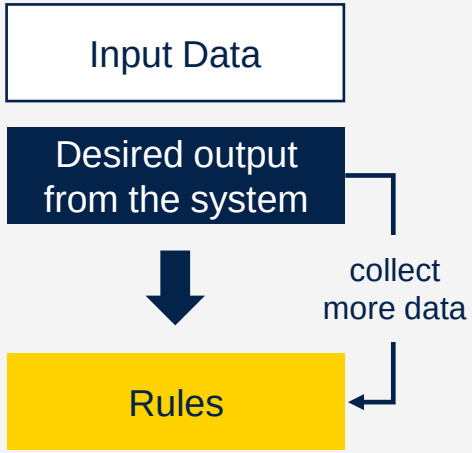


- Requires domain expertise to code
- Need to rewrite if environment evolves

Machine Learning

Rules learnt from real-world data



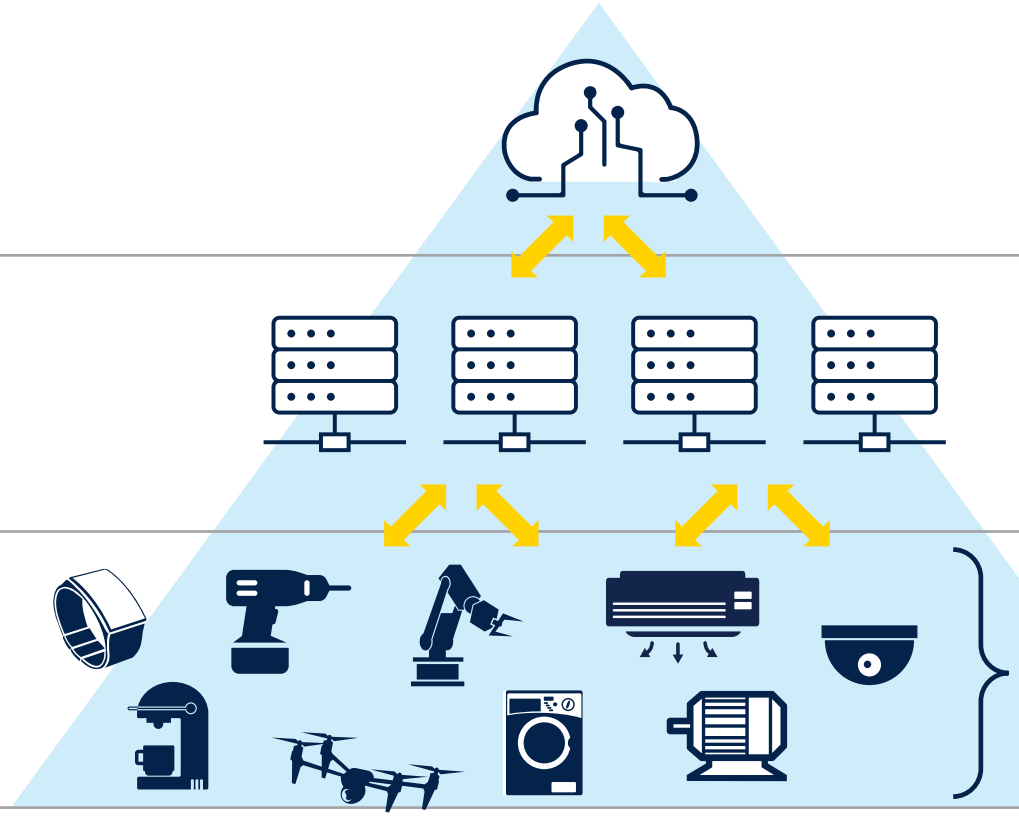


- Generate code from real-world observations
- Re-learn from data if environment evolves

life.augmented

Distributed Artificial Intelligence approach

Leverage billions of devices at the Edge!



Data Center Cloud

Analytics, storage, compute

Edge Nodes

IoT gateways, micro datacenters

Edge Things

Real time, local processing

life.augmented

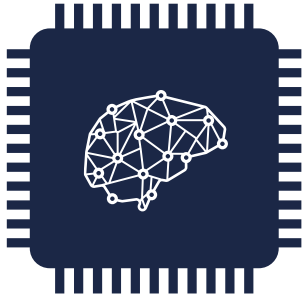
Artificial intelligence at the deep edge

Moving part of Artificial Intelligence closer to the data acquisition brings several benefits

Better user experience

No latency (real-time)

More reliability



Add new functions and services with **Embedded AI**

Optimized Cloud usage

Privacy by design (GDPR compliant)

Sustainable on energy

life.augmented

STM32 comprehensive AI ecosystem

Applications

Applicative Examples
(Function Packs)

Pre and post Processing libraries

Frameworks

NANOEDGE AI STUDIO

scikit learn

MATLAB

PyTorch

TensorFlow Lite

ONNX

Keras

STM32 Cube.AI

AI Model convertor

- Graph optimizer
- Quantizer
- Memory optimizer

Edge Hardware

STM32 series

Discovery kit

STM32 Nucleo board

Camera add-on

life.augmented

STM32 Cube.AI

life.augmented

For teams with AI expertise
Accelerate your embedded development

Train your own AI models

1

Keras
TensorFlow Lite
ONNX
PyTorch
scikit learn
MATLAB
And more…

Convert models into optimized code

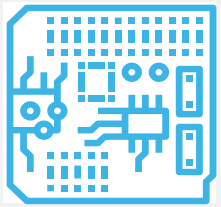
2

Neural Networks

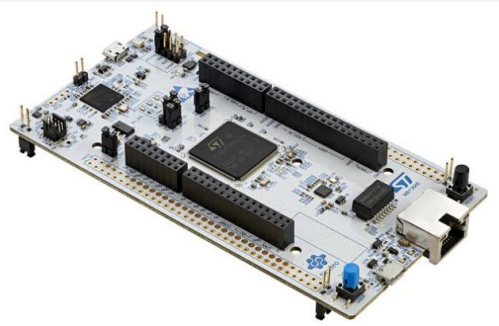
Machine Learning

Deploy on optimized runtime

3



STM32



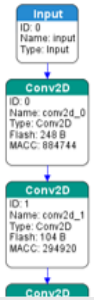
Iterate fast and increase productivity

life.augmented

The 3 pillars of STM32Cube.AI

Graph optimizer

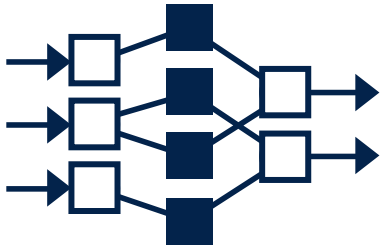
Automatically improve performance through graph simplifications & optimizations that benefit STM32 target HW architectures



- Auto graph rewrite
- Node/operator fusion
- Layout optimization
- Constant-folding…
- Operator-level info to fine-tune memory footprint and computation

Quantized model support

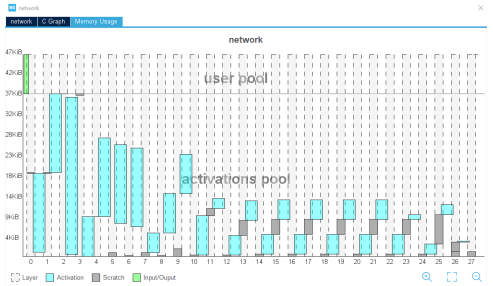
Import your quantized ANN to be compatible with STM32 embedded architectures while keeping their performance



- From FP32 to Int8
- Minimum loss of accuracy
- Code validation on target
  - Latency
  - Accuracy
  - Memory usage

Memory optimizer

Optimize memory allocation to get the best performance while respecting the constraints of your embedded design



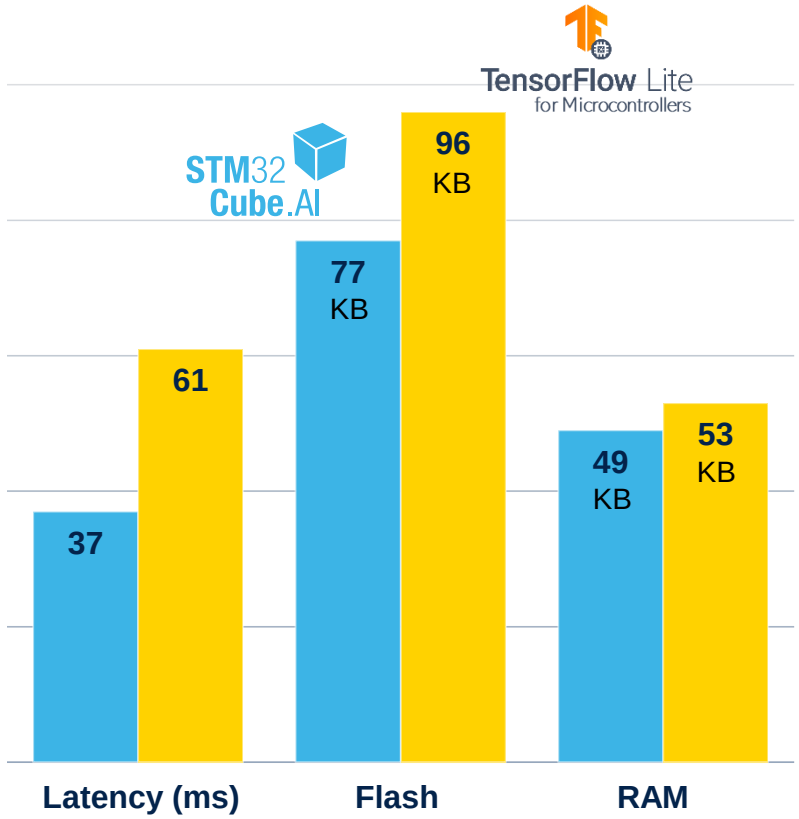
- Memory allocation
- Internal/external memory repartition
- Model-only update option

STM32Cube.AI is **free of charge**, available both in graphical interface and in command line.

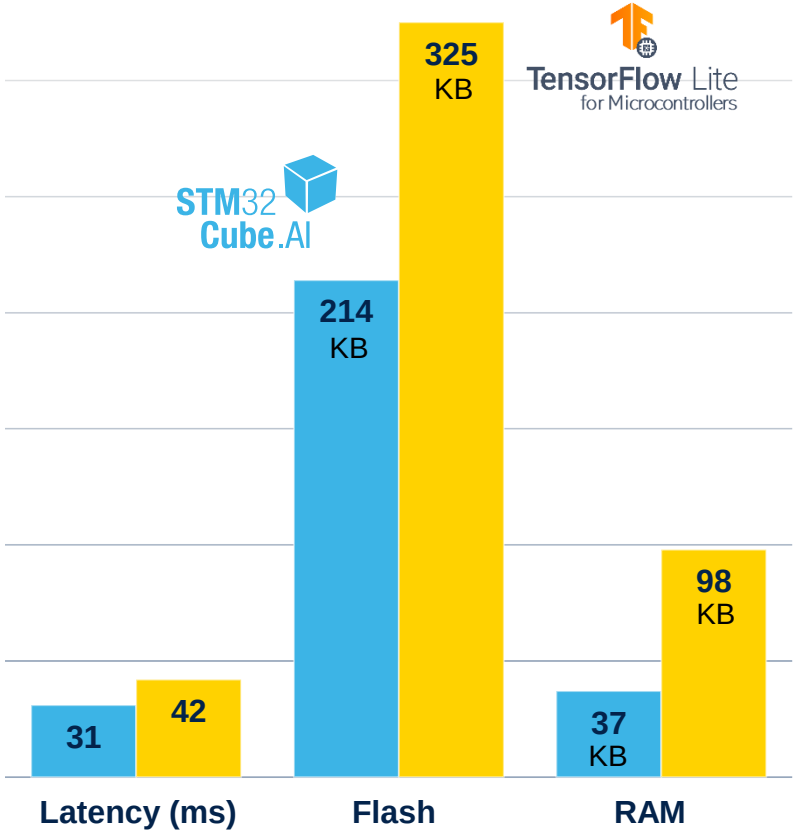
life.augmented

STM32Cube.AI
Get the best performance on STM32

Image Classif v0.5



Visual Wake Word v0.5



HW Target: STM32H723
Flash: 1Mbyte
RAM: 564 Kbytes
Freq: 550 MHz

SW Version:
X-Cube.AI v 7.0.0
TFLm v2.5.0

* the lower the better

life.augmented

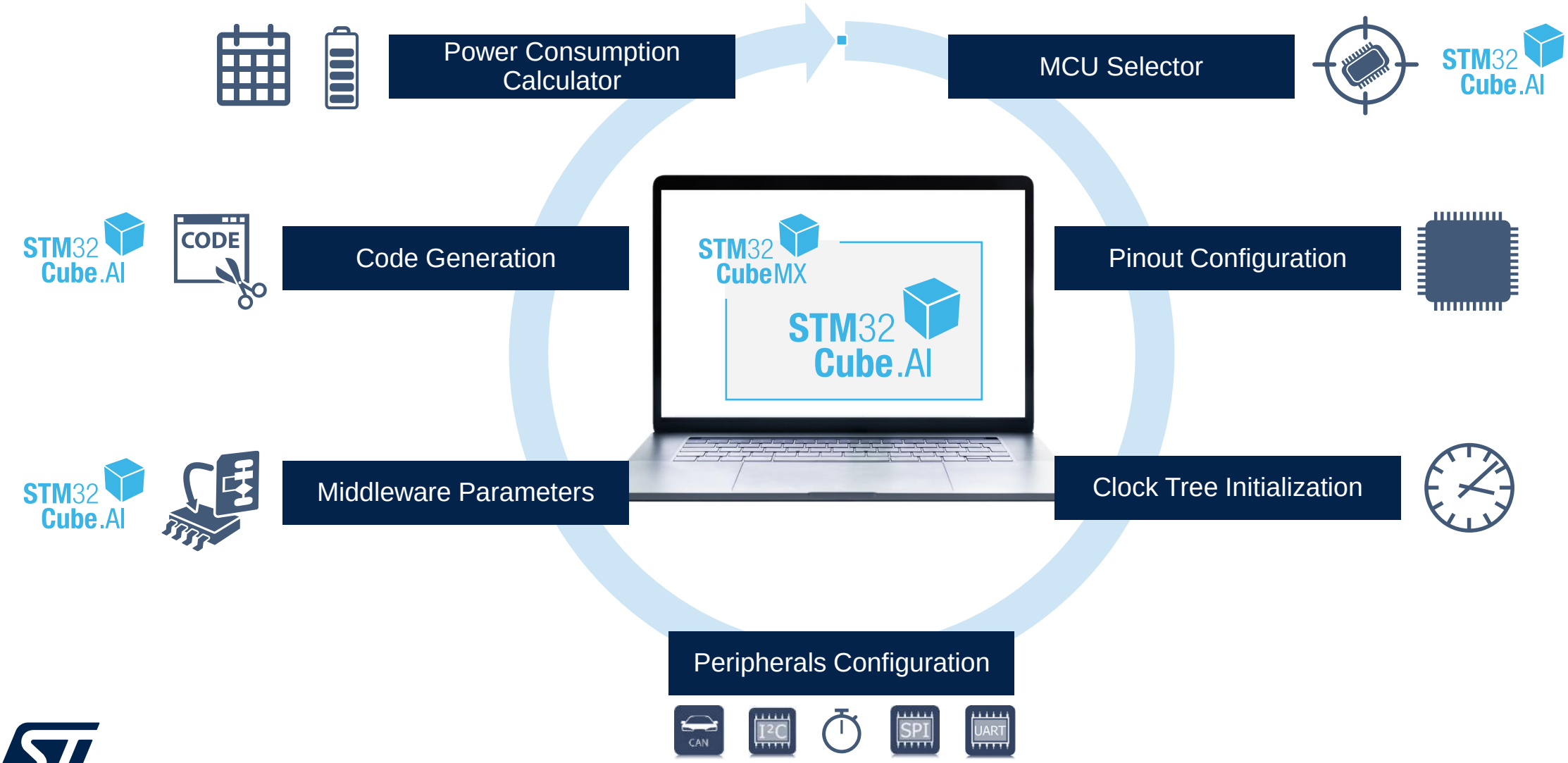
Making deep edge AI accessible to all portfolio

NANOEDGE AI STUDIO — All STM32 are compatible with **NanoEdge™Studio** ecosystem

STM32 Cube.AI — STM32 compatible with **STM32Cube.AI** ecosystem (Extension down to M0 since 7.1.0 version)

	-M0	-M0+	-M3	-M33	-M4	-M7	dual -A7& -M4
Arm® Cortex® core							
MPU							STM32**MP1** 4158 CoreMark 800 MHz Cortex -A7 209 MHz Cortex -M4
High Perf MCUs			STM32**F2** 398 CoreMark 120 MHz		STM32**F4** 608 CoreMark 180 MHz	STM32**H7** 3224 CoreMark 240 MHz Cortex -M4 480 MHz Cortex -M7	STM32**F7** 1082 CoreMark 216 MHz
Mainstream MCUs	STM32**F0** 106 CoreMark 48 MHz	STM32**G0** 142 CoreMark 64 MHz	STM32**F1** 177 CoreMark 72 MHz		STM32**F3** 245 CoreMark 72 MHz	STM32**G4** 550 CoreMark 170 MHz	
Ultra-low Power MCUs		STM32**L0** 75 CoreMark 32 MHz	STM32**L1** 93 CoreMark 32 MHz	STM32**L5** 424 CoreMark 110 MHz STM32U**5** 651 CoreMark 160 MHz	STM32**L4** 273 CoreMark 80 MHz	STM32**L4+** 409 CoreMark 120 MHz	
Wireless MCUs					STM32**WL** 161 CoreMark 48 MHz	STM32**WB** 216 CoreMark 64 MHz	

STM32Cube.AI
The STM32CubeMX expansion pack for ML



life.augmented

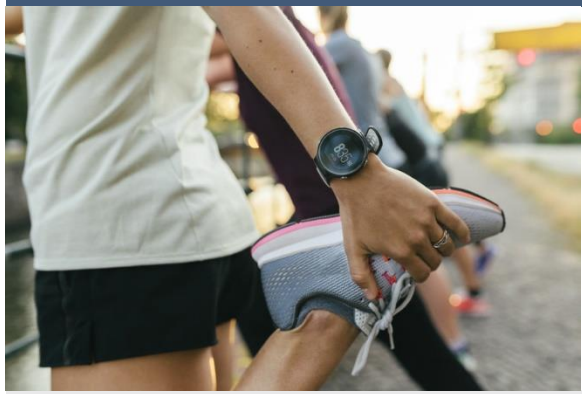
Integrate your ML models more easily with our application-oriented code examples

Time series-based monitoring

FP-AI-MONITOR1

- Predictive maintenance and much more sensor-monitoring apps
- Runs Libraries from NanoEdge AI Studio

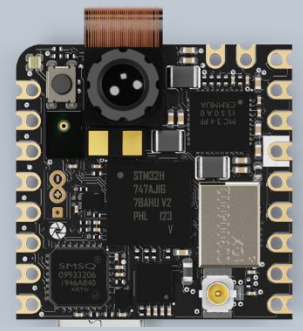
Audio and sensing



FP-AI-SENSING1

- Human Activity Recognition
- Acoustic Scene Classification
- Data logging, labeling and result on BLE applications

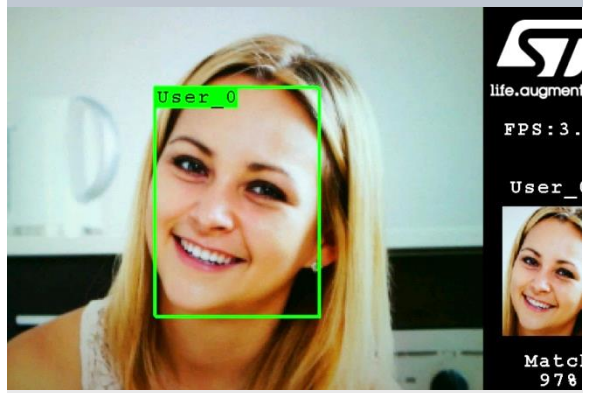
Computer vision



FP-AI-VISION1

- Food recognition (CNN)
- Person presence detection (CNN)
- People counting (Object detection NN)
- Image processing Library

Face recognition



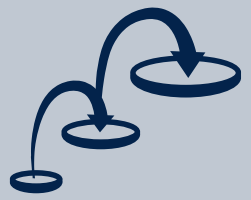
FP-AI-FACEREC1

- Face detection and recognition
- Fully functional without cloud connection

life.augmented

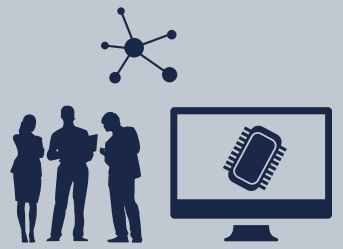
We provide everything to kick off your project

Design documentation



Getting started

Be guided step-by-step to learn STM32 ecosystem

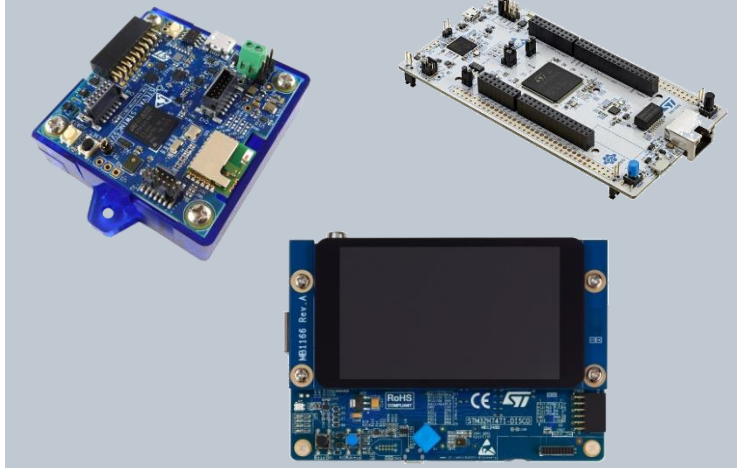


Development zone

Get started on application development and project sharing

- **Wiki by ST** is a great forum to learn and start developing AI on STM32!
- Videos of application examples
- Massive Open Online Course (MOOC)

Hardware and software tools



- Evaluation platforms for STM32 MCU/MPU
- Extra sensor boards
- Full software suite

Support & Updates

- **ST Community**: STM32 ML & AI group
- Distributor certified FAE
- Support center
- Newsletter

life.augmented

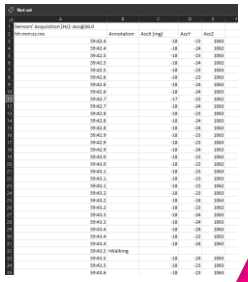
STM32Cube.AI Deeper dive

life.augmented

The key steps behind Neural Networks

Neural Network (NN) Model Creation

Operating Mode



Capture data

1

Clean, label data
Build NN topology

2

Train NN Model

3

Convert NN into
optimized code for MCU

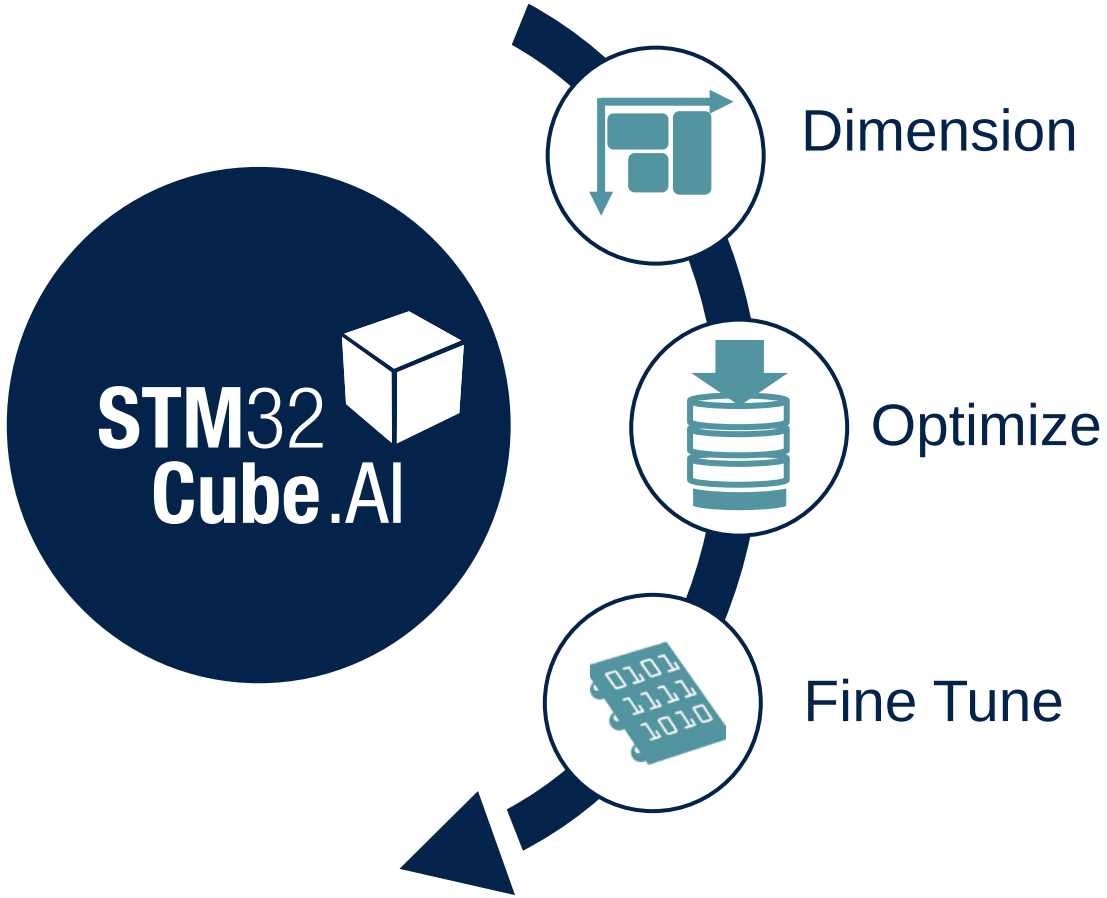
4

Process & analyze new
data using trained NN

5

STM32Cube.AI main features

STM32Cube.AI is available both as graphical and command line interface



- ✓ Quickly assess model footprint requirements
- ✓ Select and configure MCU in STM32CubeMX
- ✓ Review model layers in STM32Cube.AI

- ✓ Generate C-code for pre-trained model
- ✓ Support quantized models to reduce RAM, flash and latency with minimal loss of accuracy
- ✓ Use light run-time libraries
- ✓ Optimize for performance

- ✓ Optimize memory allocation
- ✓ Fine control of weight mapping
- ✓ Split between internal and external memory
- ✓ Update model without full FW update

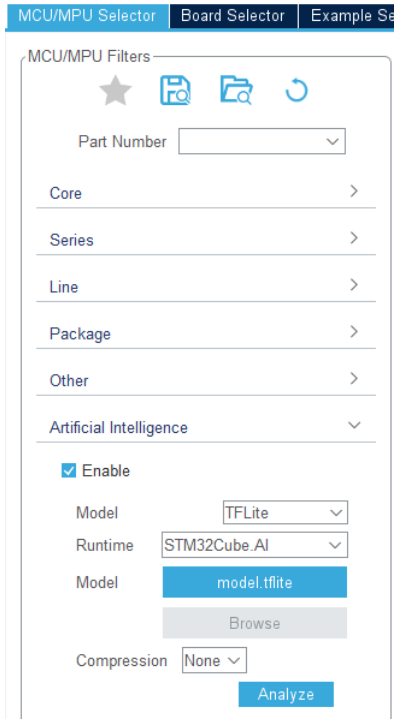
And quickly iterate thanks to on-target validation

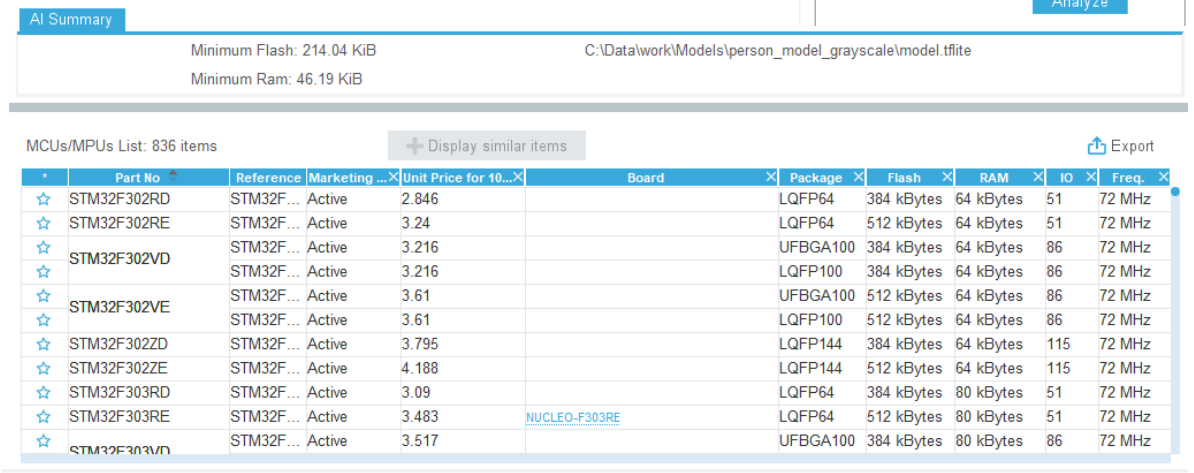
life.augmented

STM32Cube.AI user flow (1/3)

1. Select MCU
2. Optimize and validate
3. Generate project and integrate

- Load model and select an AI runtime
- Analyze minimal required footprint
- Select corresponding STM32 MCUs

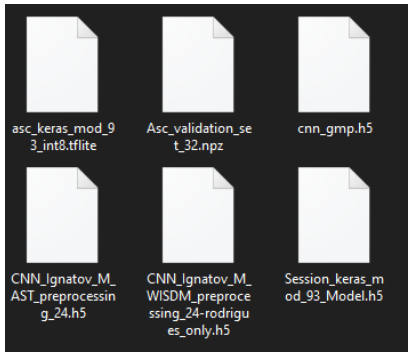




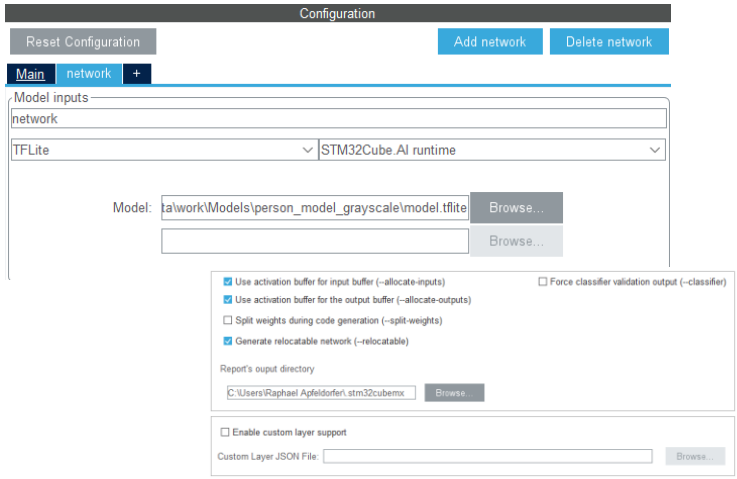
Part No	Reference	Marketing ...	Unit Price for 10...	Board	Package	Flash	RAM	IO	Freq.
STM32F302RD	STM32F...	Active	2.846		LQFP64	384 kBytes	64 kBytes	51	72 MHz
STM32F302RE	STM32F...	Active	3.24		LQFP64	512 kBytes	64 kBytes	51	72 MHz
STM32F302VD	STM32F...	Active	3.216		UFBGA100	384 kBytes	64 kBytes	86	72 MHz
	STM32F...	Active	3.216		LQFP100	384 kBytes	64 kBytes	86	72 MHz
STM32F302VE	STM32F...	Active	3.61		UFBGA100	512 kBytes	64 kBytes	86	72 MHz
	STM32F...	Active	3.61		LQFP100	512 kBytes	64 kBytes	86	72 MHz
STM32F302ZD	STM32F...	Active	3.795		LQFP144	384 kBytes	64 kBytes	115	72 MHz
STM32F302ZE	STM32F...	Active	4.188		LQFP144	512 kBytes	64 kBytes	115	72 MHz
STM32F303RD	STM32F...	Active	3.09		LQFP64	384 kBytes	80 kBytes	51	72 MHz
STM32F303RE	STM32F...	Active	3.483	NUCLEO-F303RE	LQFP64	512 kBytes	80 kBytes	51	72 MHz
STM32F303VD	STM32F...	Active	3.517		UFBGA100	384 kBytes	80 kBytes	86	72 MHz

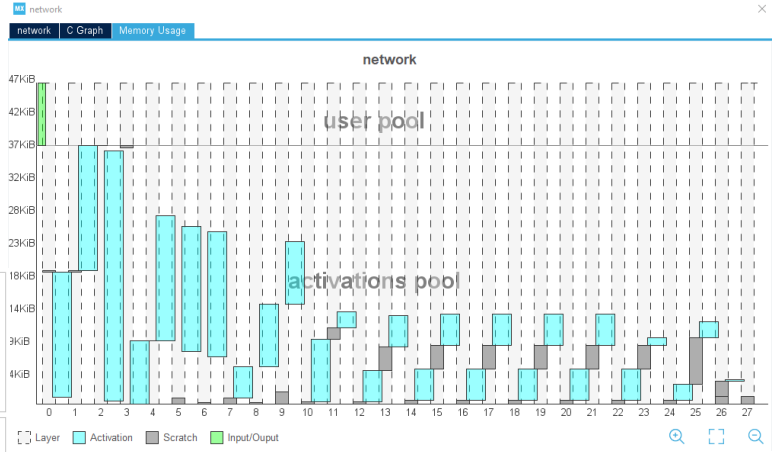
STM32Cube.AI user flow (2/3)

1. Select MCU
2. Optimize and validate
3. Generate project and integrate



- Model complexity and footprint analysis
- Fine tune memory allocation with optimizations and GUI
- Optimize system parameters and clock tree
- Extend model with your own customer layers

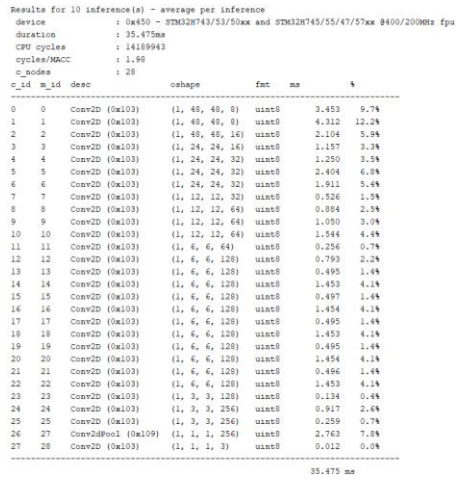




- Validate on desktop with your own dataset
- Validate on target and check inference time

Validate on desktop

Validate on target

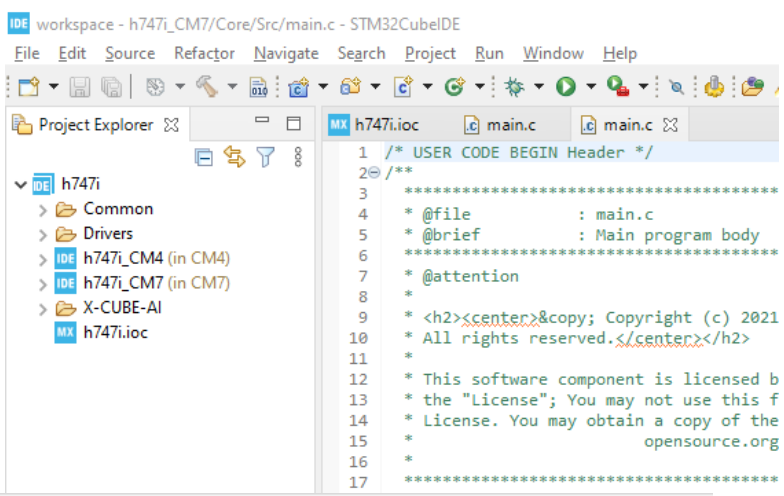


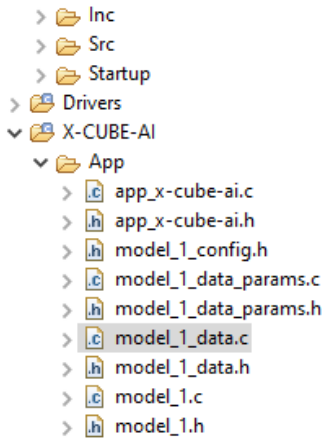
STM32Cube.AI user flow (3/3)

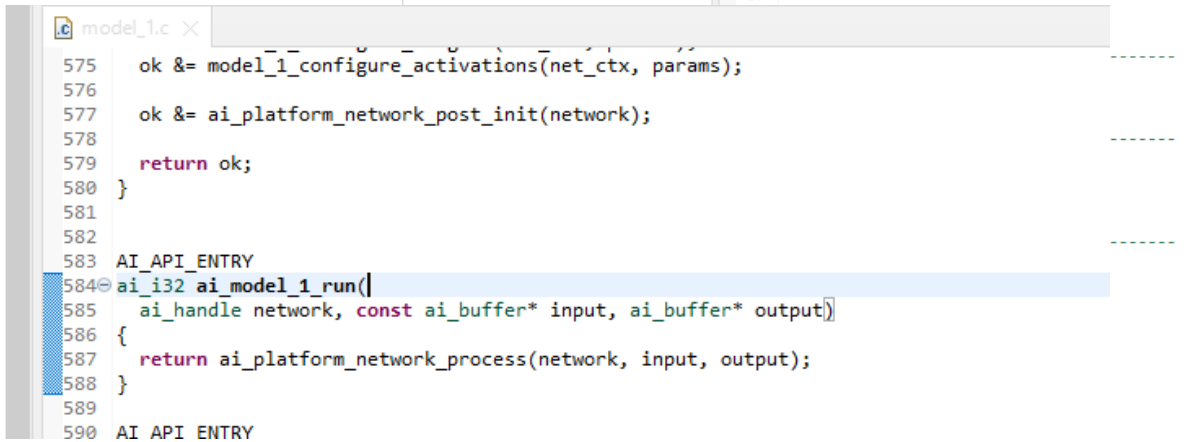
1. Select MCU
2. Optimize and validate
3. Generate project and integrate

GENERATE CODE

- Generate Application Template
- Integrate with your application-specific code in your favorite IDE
- Perform system tests



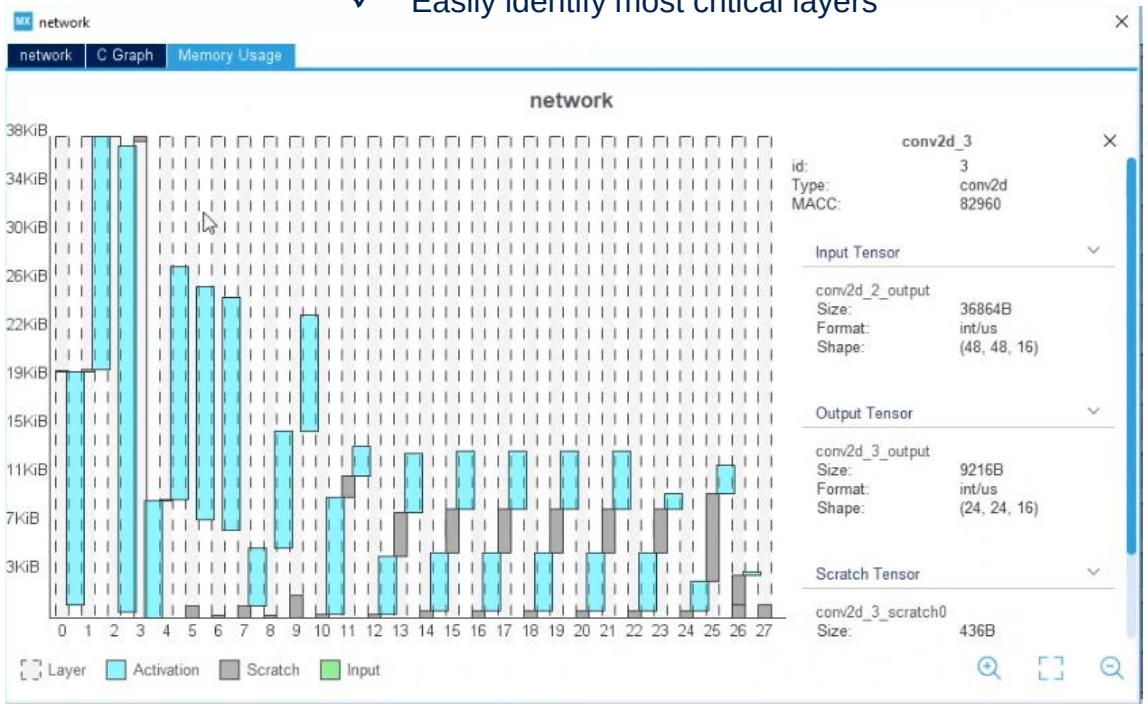




Fine-tuning memory allocation

Model RAM consumption can be displayed layer per layer

✓ Easily identify most critical layers

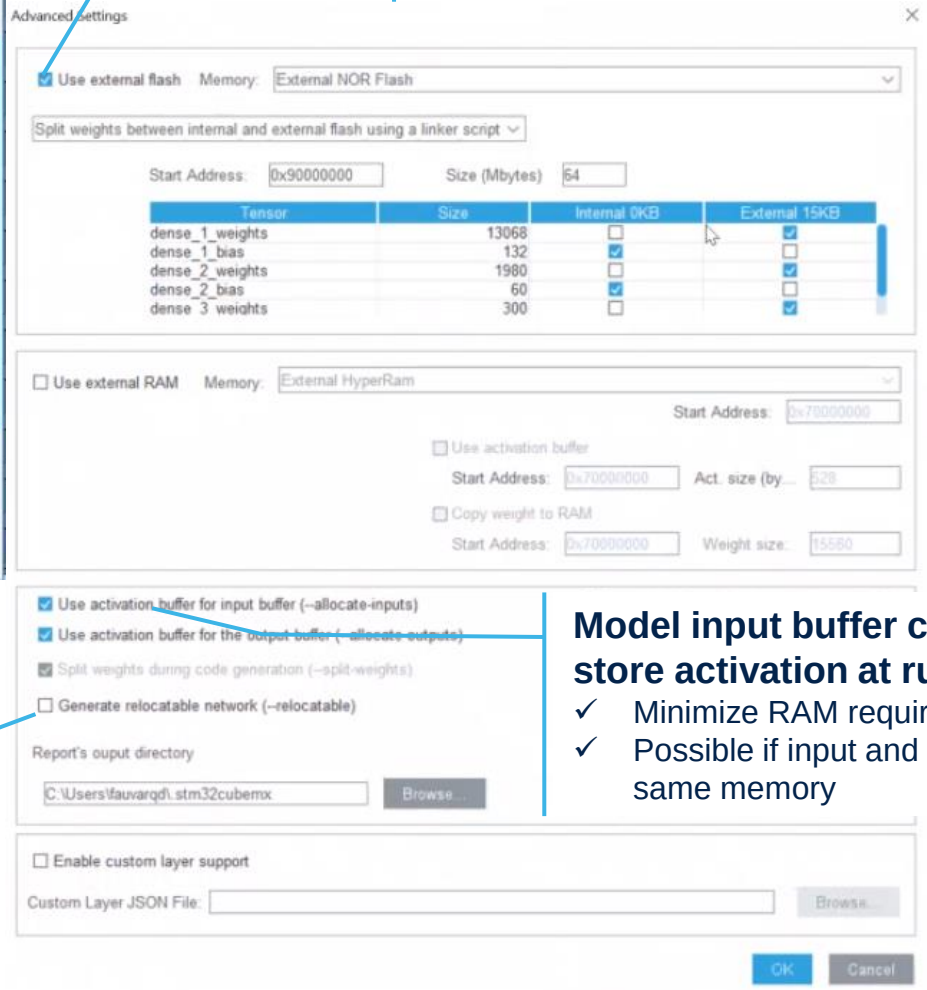


Model memory allocation is easily controlled

✓ Make use of external memory based on model requirements and application footprint

✓ Mapped in non-contiguous internal flash section

✓ Mapping partially in internal and external flash memories



Model input buffer can be re-used to store activation at run-time

✓ Minimize RAM requirements

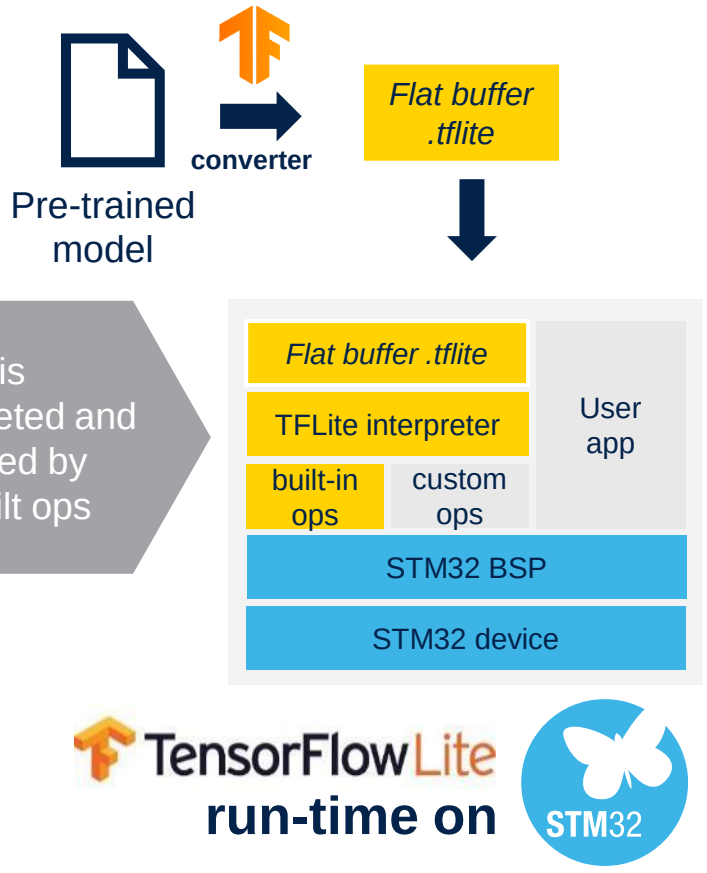
✓ Possible if input and activation buffers are in same memory

Enable relocatable network

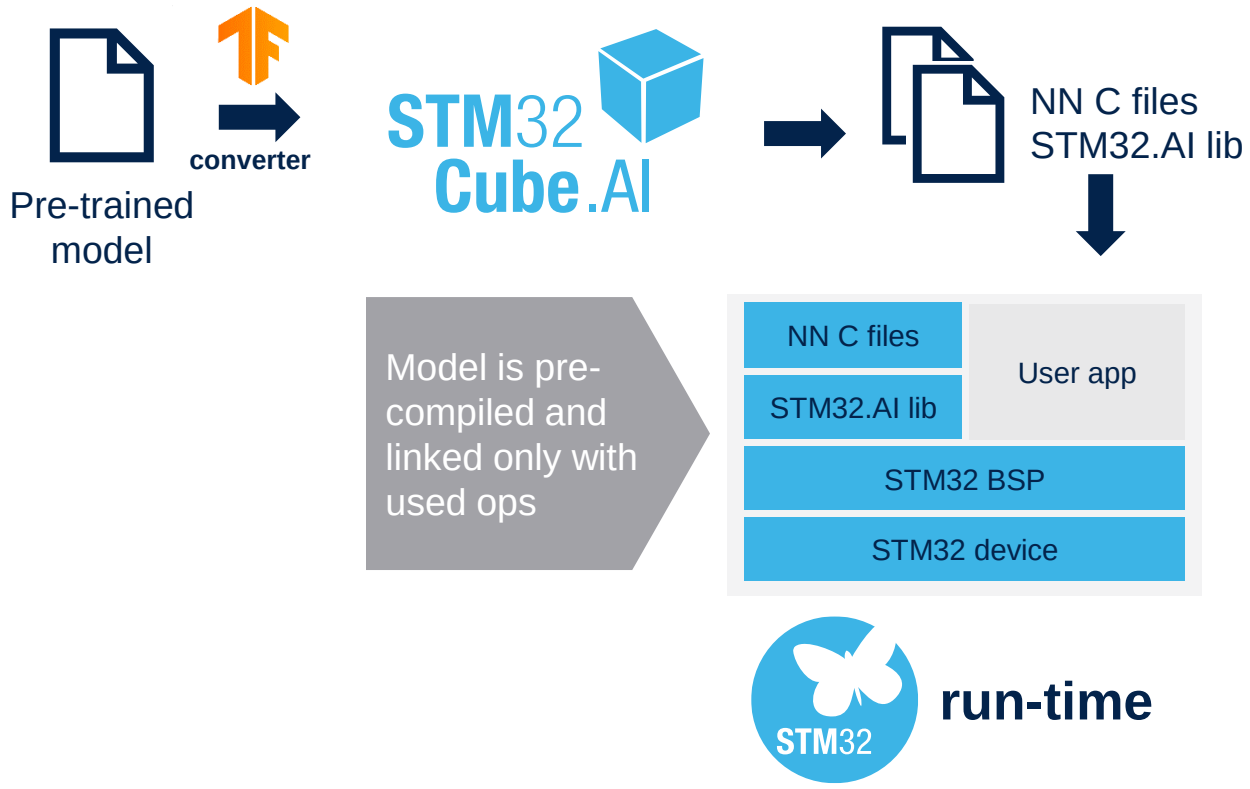
✓ A separate binary is generated for the library and the network to enable model upgrade

Possible conversion strategies: Network code generation and interpreter

More Flexible: TensorFlow Lite interpreter mode



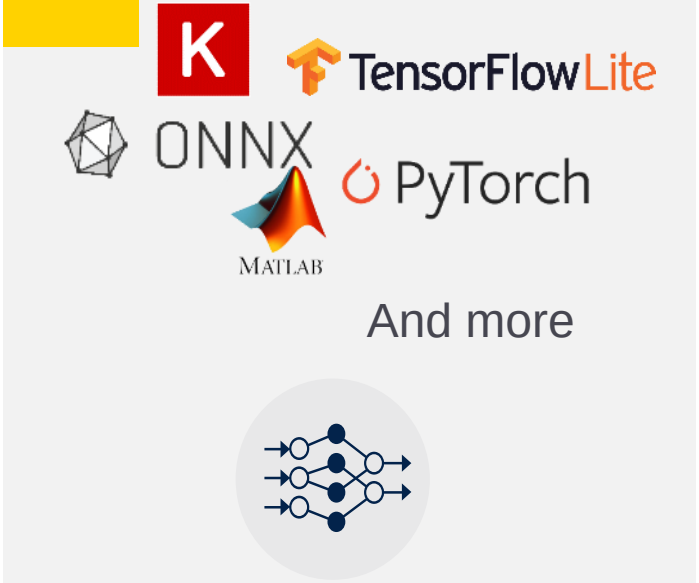
More optimized: Optimized C code generated by STM32Cube.AI



STM32Cube.AI with ML extension

Leverage NN and ML in a unique tool optimized for STM32

Train Neural Network using any major AI frameworks



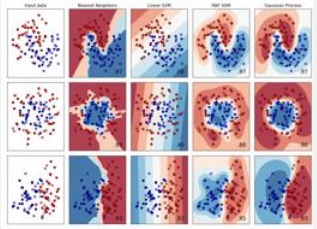
Train ML model using SciKit Learn

- Efficiently develop your ML solution on STM32
- Add ML layers to complement Neural Network model
- Implement pre-processing steps based on ML

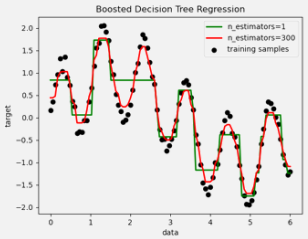
STM32Cube.AI with ML extension

STM32 Cube.AI + scikit learn

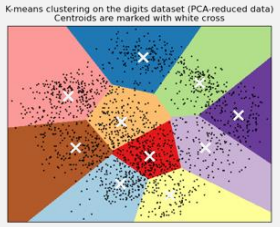
Classification



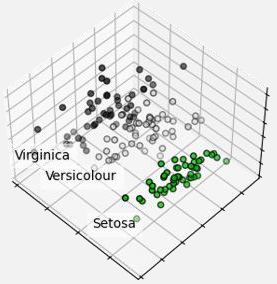
Regression



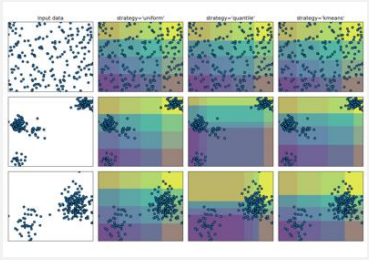
Clustering



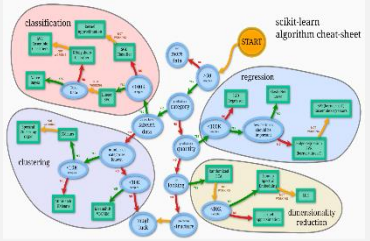
Dimensionality reduction



Preprocessing



Model selection



life.augmented

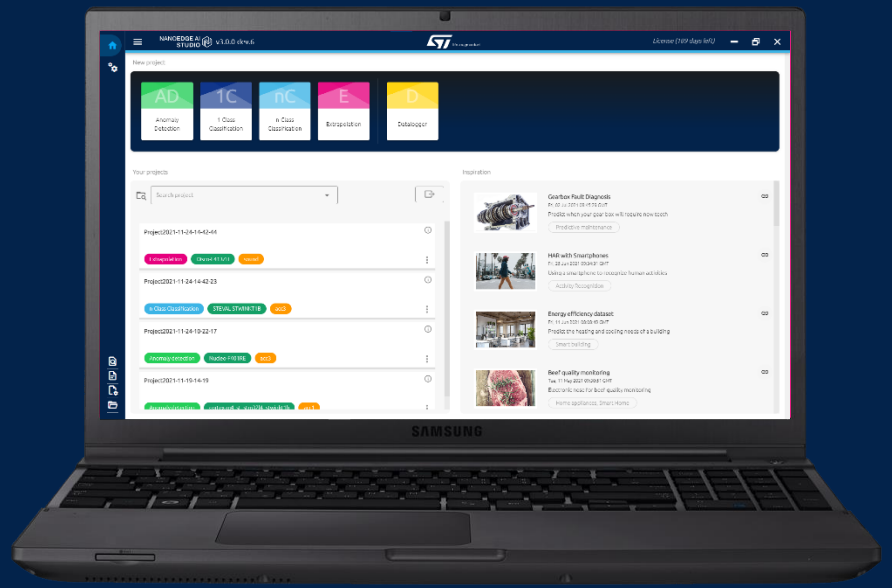
NanoEdge AI Studio

life.augmented

NanoEdge AI Studio V3
Same easiness, more power

ON THE PC

1 Create the library, ONCE.

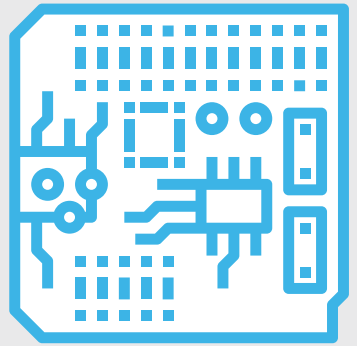


Standalone PC (Win/Linux) solution

Create and embed a self learning engine

ON THE MCU

2 Use the library, MANY TIMES.

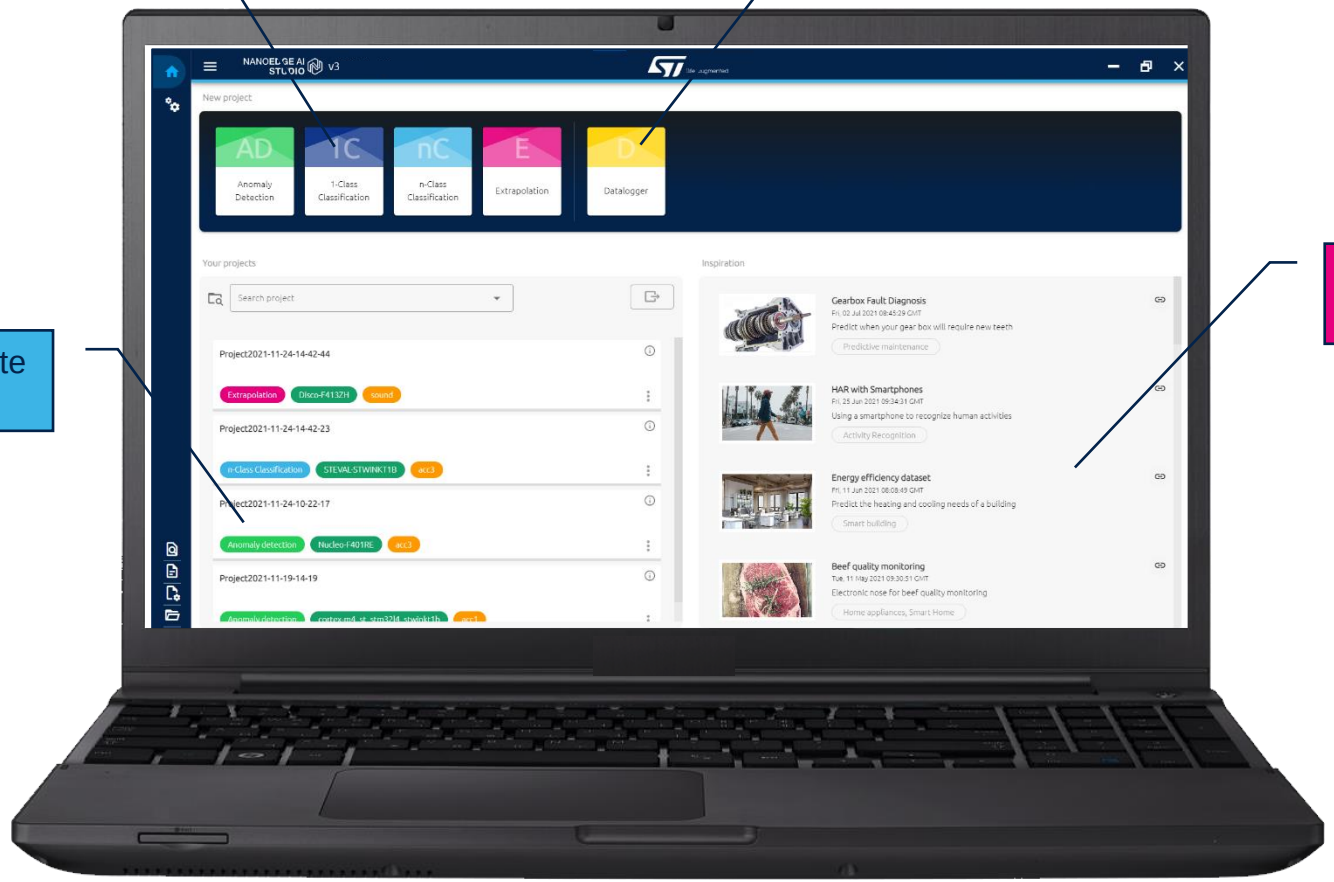


For anomaly detection, the model is self-trained at the Edge.

New user interface, more functions, better user experience

New families of Machine Learning algorithms

New Datalogging experience



Easily retrieve or create projects

Get inspired by Multiple uses cases

life.augmented

NANOEDGE AI STUDIO

Our customers have increasingly ambitious use cases for ever smarter products

AD
Anomaly Detection

1C
1-Class Classification

nC
n-Class Classification

E
Extrapolation

"My pumps are installed in a variety of environments that I can't anticipate.
I want them to autonomously adapt to their target environment and detect anomalies by themselves."

"I know exactly how my pumps behave.
I want to detect any outliers."

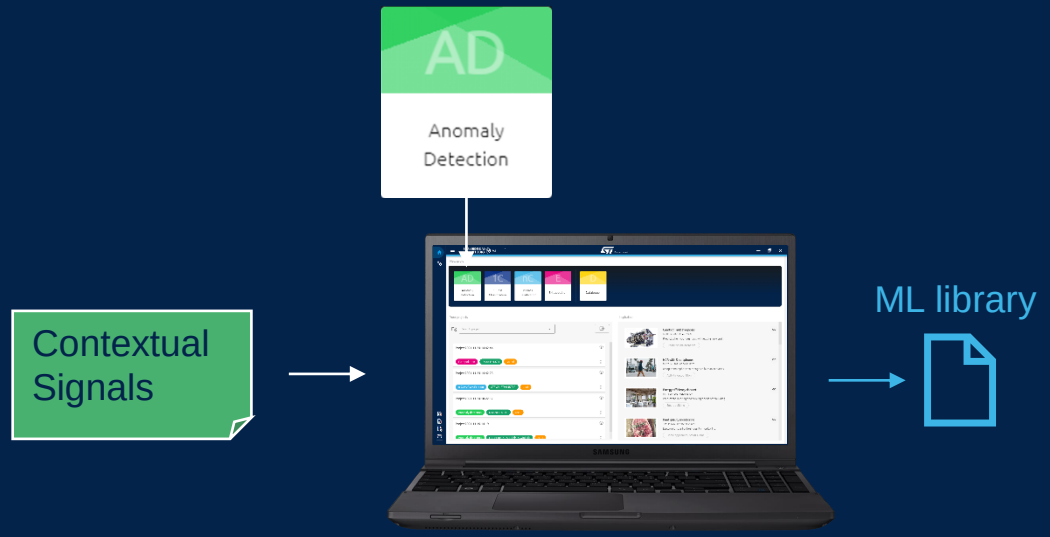
"I know the signals when a pump is experiencing, for example, ball bearing or cavitation problems.
I want to know by name what problems are occurring.“

“I know several vibration values of my machine.
I want to anticipate when a specific vibration level will be reached so that I have time to take corrective actions before reaching that limit.”

Anomaly detection use-case

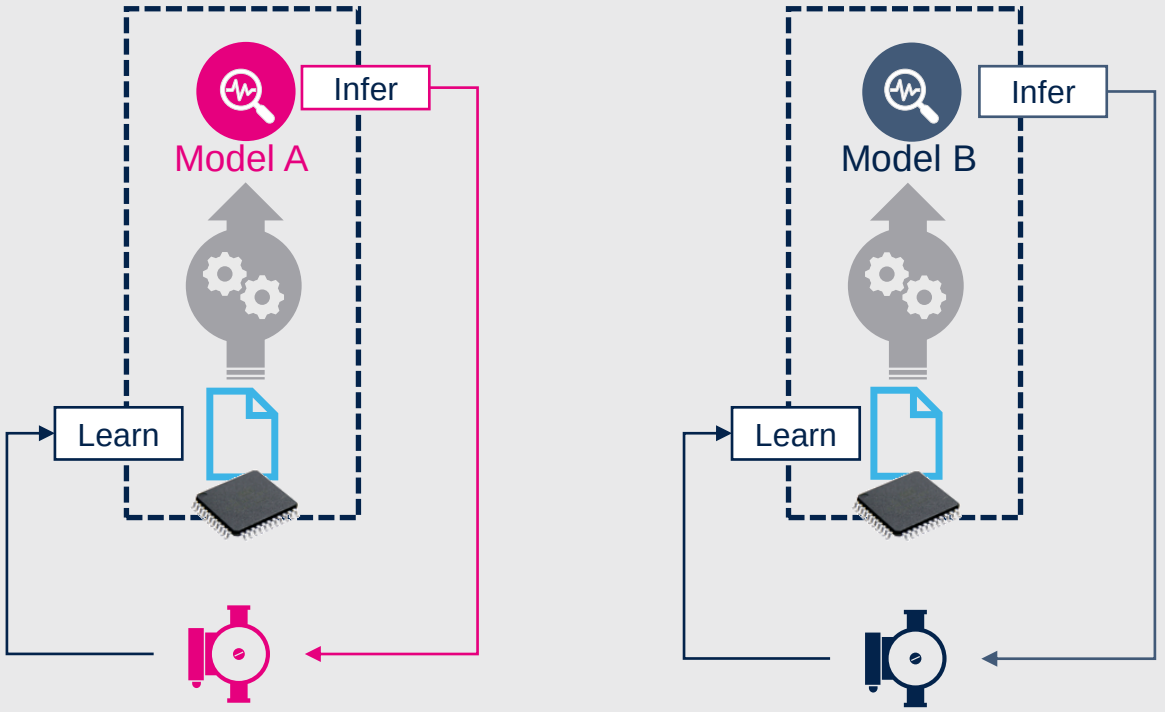
ON THE PC

1 Creation of an **ANOMALY DETECTION** Machine Learning library



ON THE MCU

2 Use of an **ANOMALY DETECTION** Machine Learning library



NEW

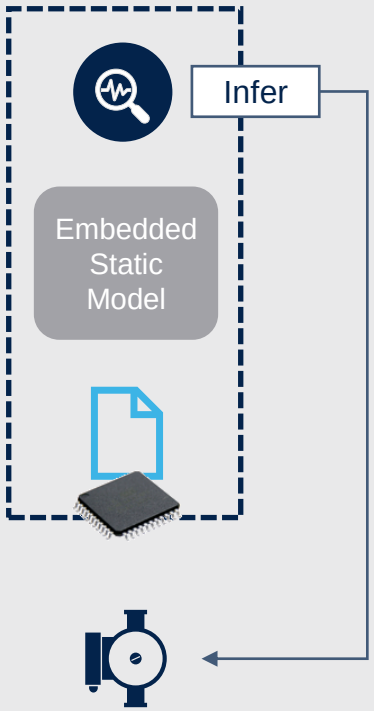
One class classification use-case

ON THE PC

1 Creation of a **ONE CLASS CLASSIFICATION** Machine Learning library

ON THE MCU

2 Use of an **ONE CLASS CLASSIFICATION** Machine Learning library



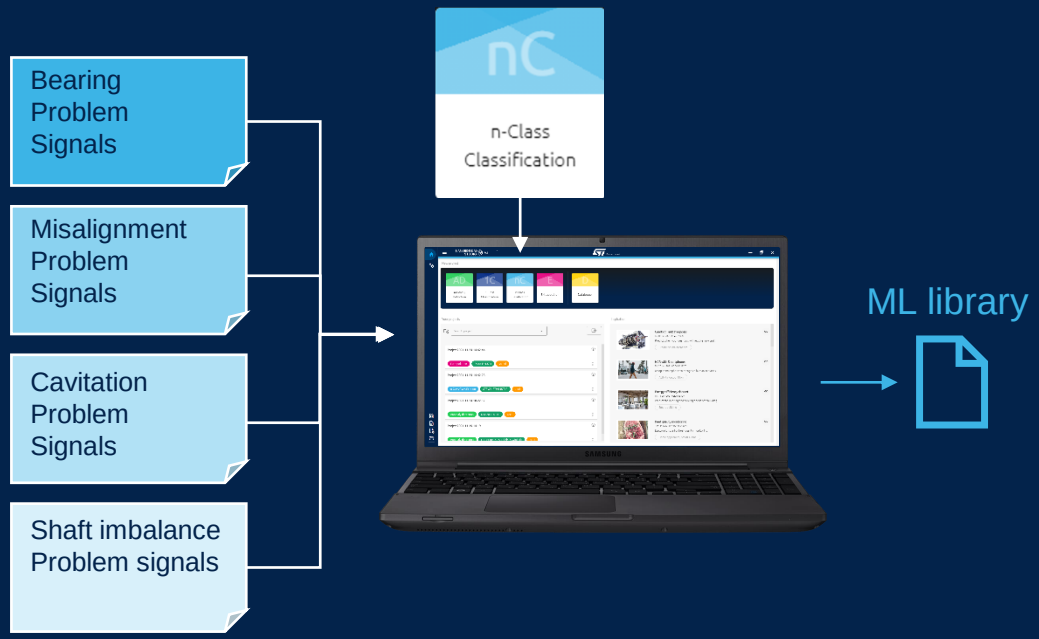
life.augmented

NANOEDGE AI STUDIO

n Class classification use-case

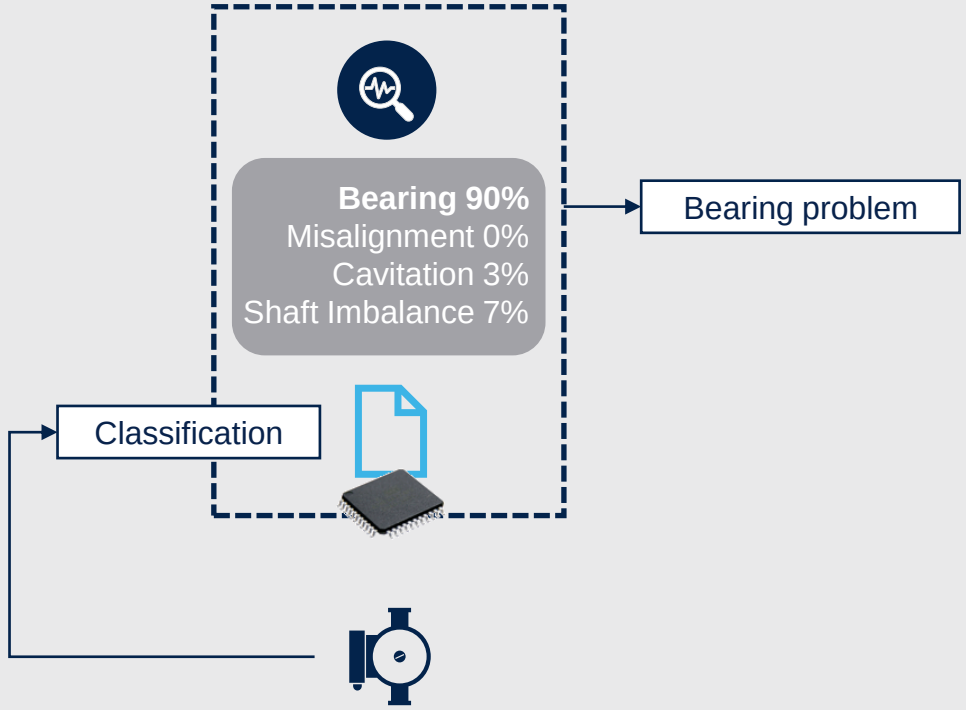
ON THE PC

1 Creation of a **n CLASS CLASSIFICATION** Machine Learning library



ON THE MCU

2 Use of an **n CLASS CLASSIFICATION** Machine Learning library



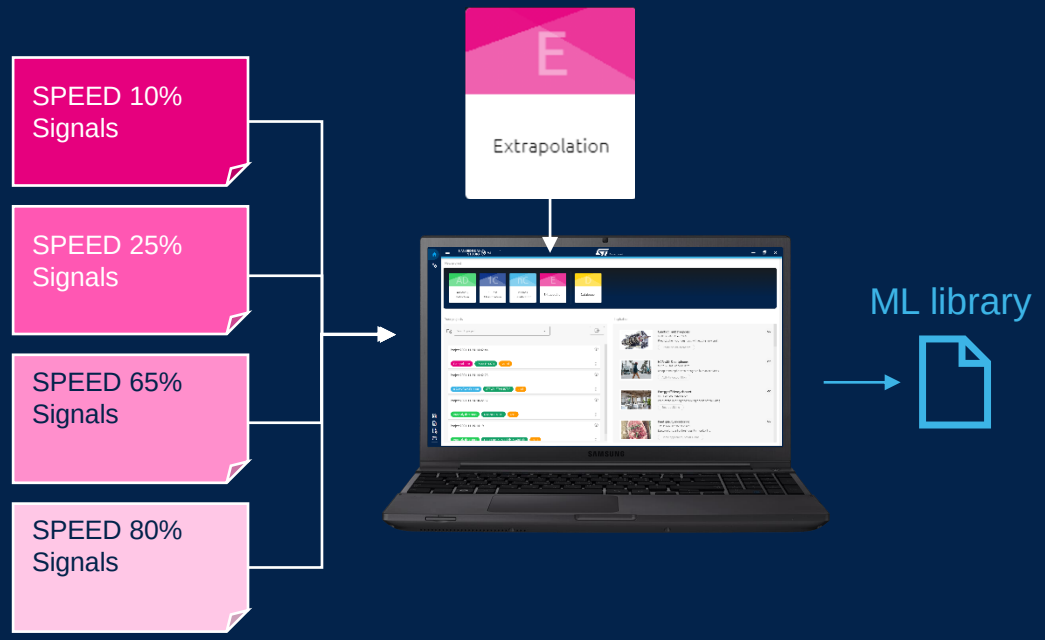
life.augmented | NANOEDGE AI STUDIO

NEW

Extrapolation use-case

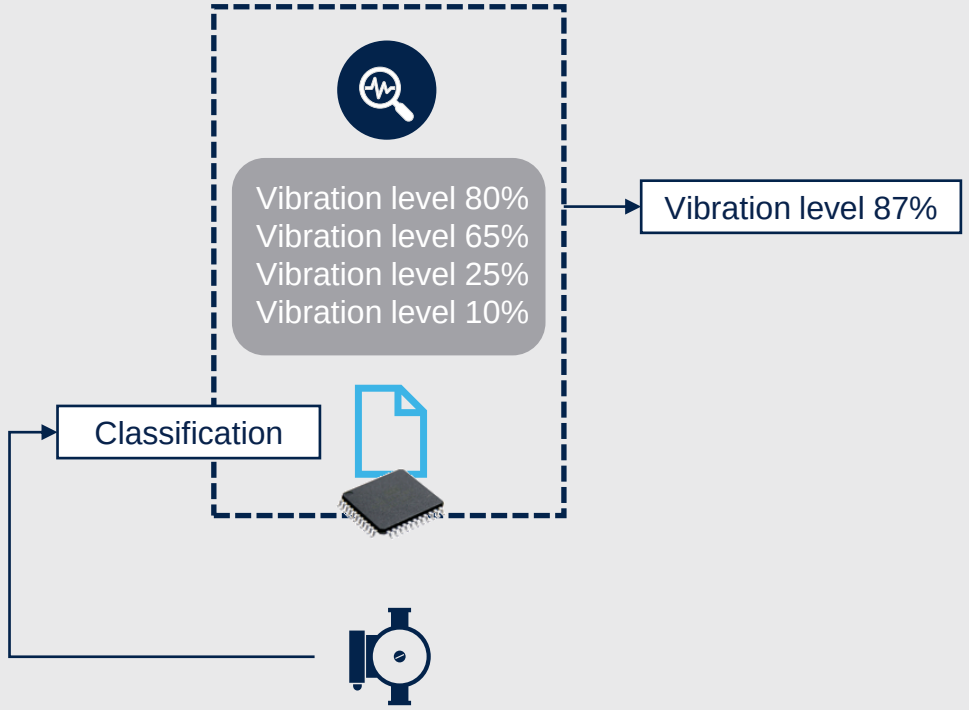
ON THE PC

1 Creation of an **EXTRAPOLATION** Machine Learning library



ON THE MCU

2 Use of an **EXTRAPOLATION** Machine Learning library

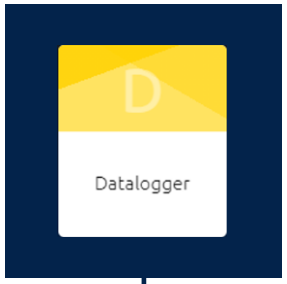


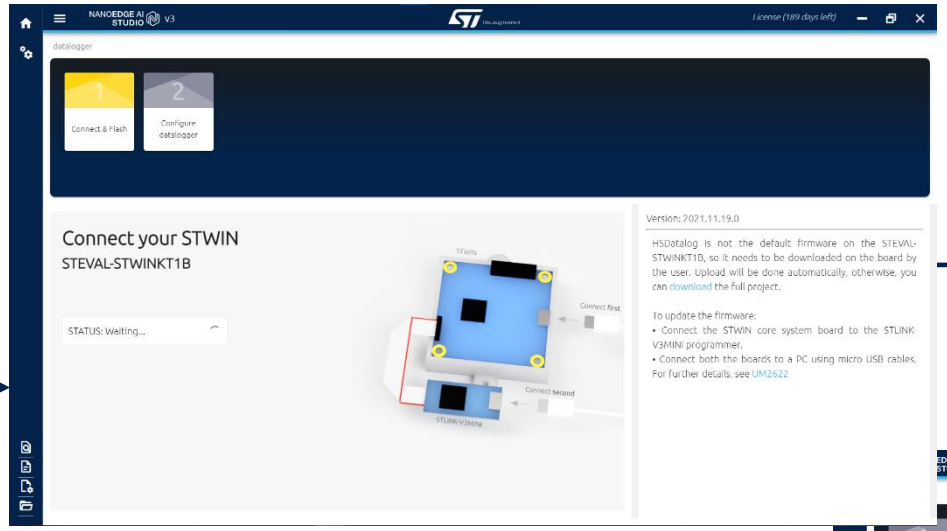
life.augmented

NANOEDGE AI STUDIO

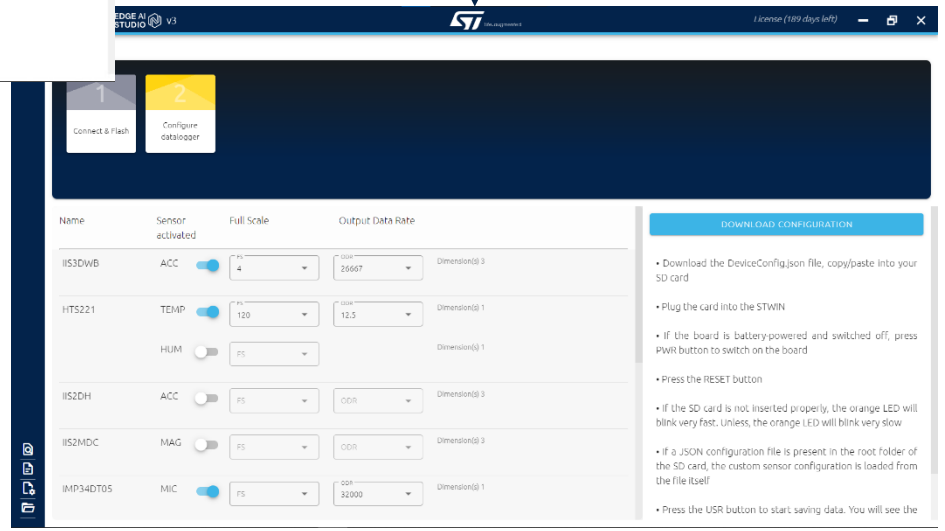
NEW

From idea to datalogging in a matter of minutes



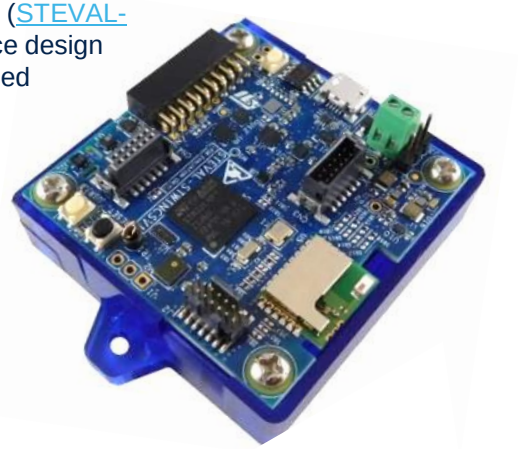


- Streamlined data logging process
- No code
- All settings done using a graphic interface



The STWIN SensorTile wireless industrial node (STEVAL-STWINKT1B) is a development kit and reference design that simplifies prototyping and testing of advanced industrial IoT applications such as condition monitoring and predictive maintenance

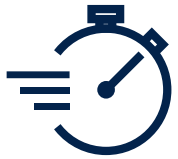
The kit features a core system board with a range of embedded industrial-grade sensors and an ultra-low-power microcontroller



life.augmented

NANOEDGE AI STUDIO

How NanoEdge™ can grow your business



Save time, resources & money

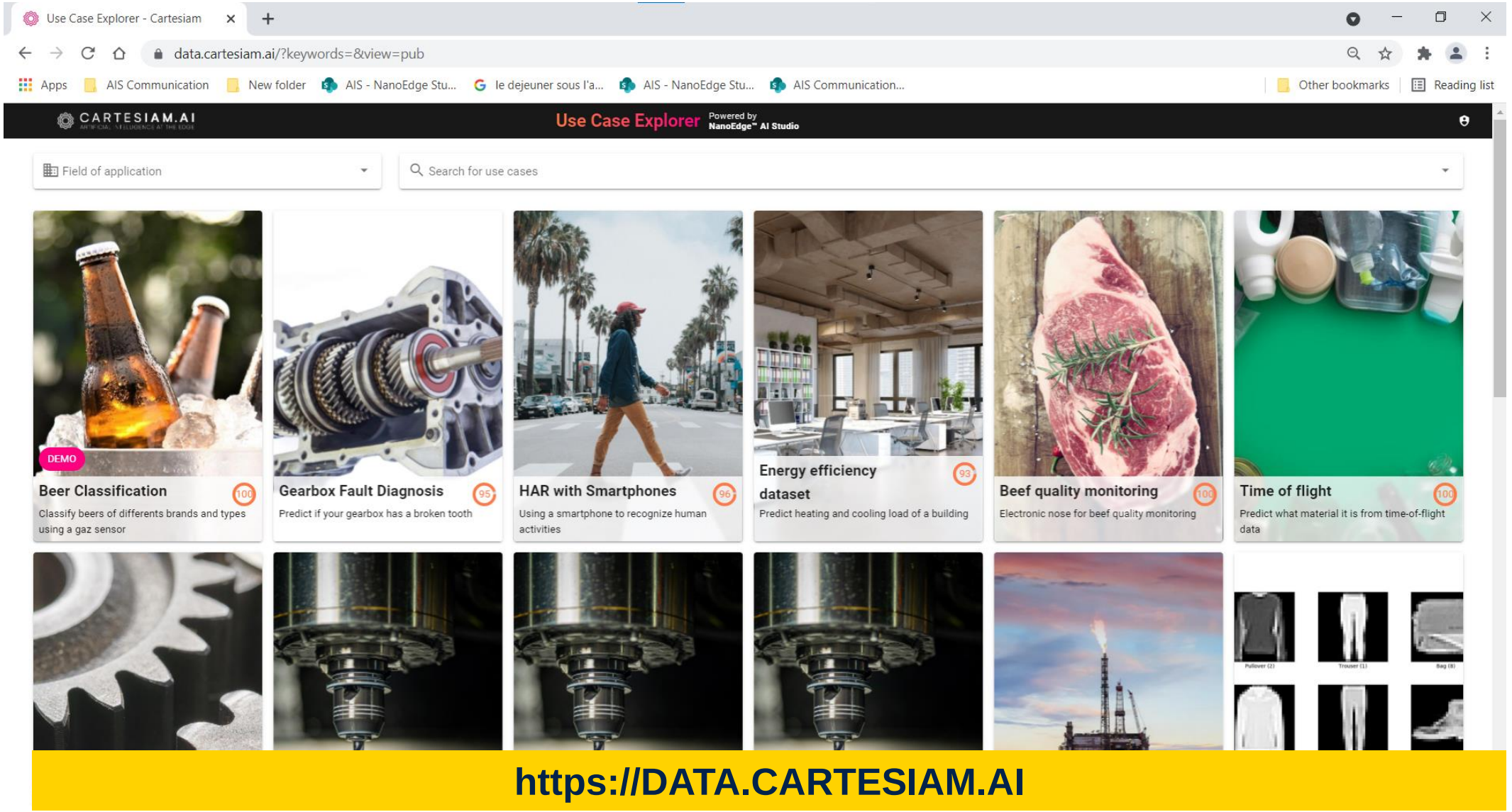
- Full AI prototype for **less than 3-month data scientist salary**
- No AI experts required
- Works on M0+
- Fast time to market: your AI solution **in a few weeks**

Raise customer satisfaction & loyalty

- Set yourself apart from competition
- Position your company as **an innovative business**
- **Improve** level of service or **offer** a new service

life.augmented

And many more



https://DATA.CARTESIAM.AI

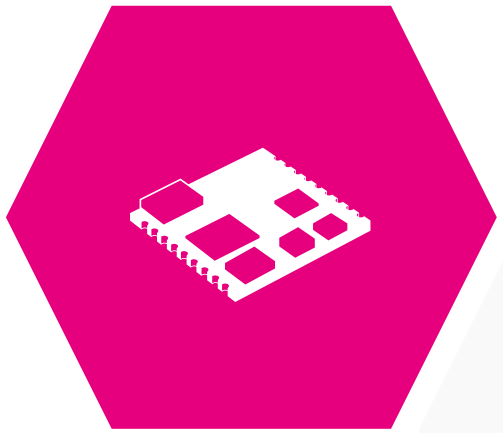
STM32 EVK Board for AI

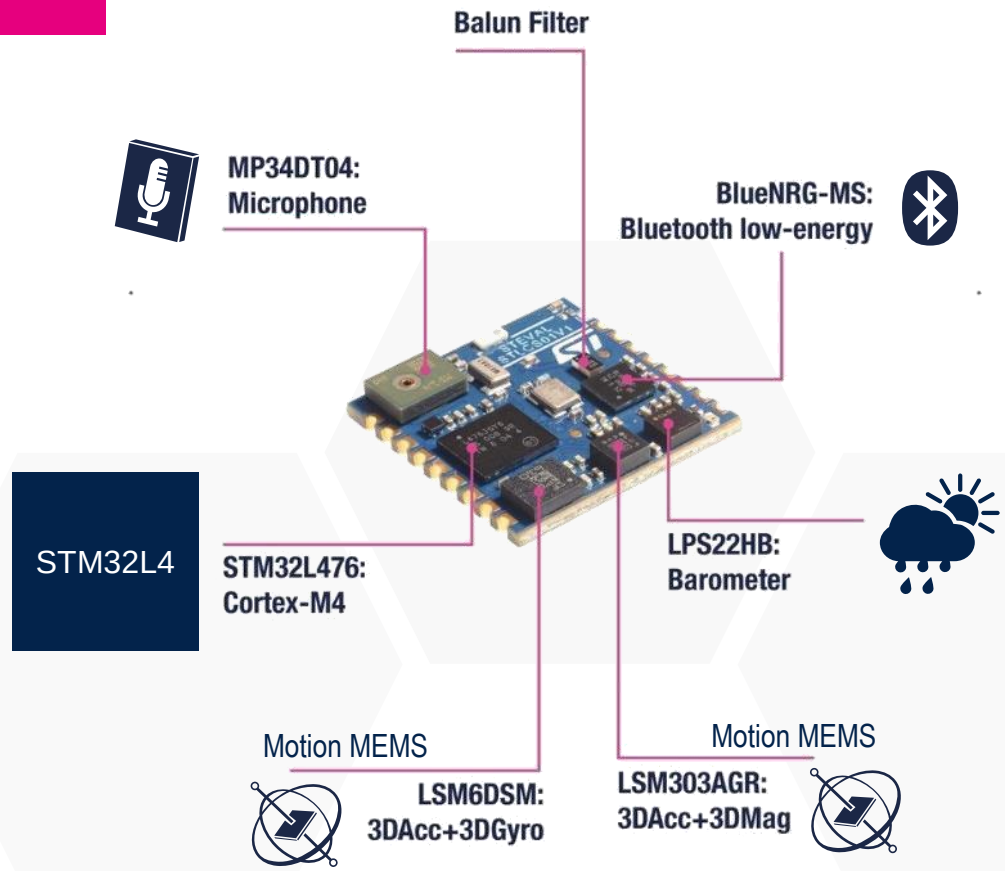
life.augmented

Example form factor hardware to capture and process data

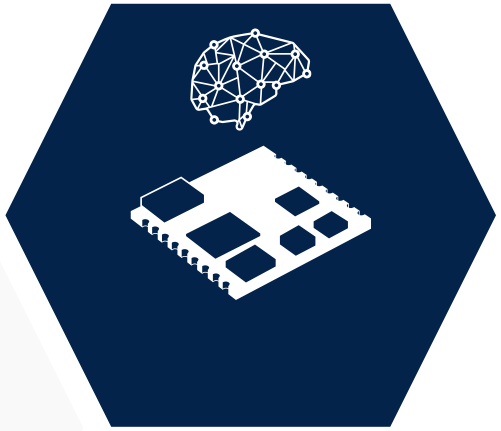
SensorTile

Capture Data





Inference on STM32L476



www.st.com/SensorTile
www.st.com/SensorTile-edu

Fast go to market module to capture data with more accuracy

SensorTile.Box

Capture Data

Inference on STM32L4R9

Microsoft IoT Services ready

Microsoft Azure

More advanced, high accuracy and low power sensors

- First Inertial module with Machine Learning capabilities.
- Motion (accelerometer and gyroscope, magnetometer) and slow motion (inclinometer)
- Altitude (pressure), environment (pressure, temperature, humidity, compass) and sound (sound and ultrasound analog microphone)
- Microsoft IoT services ready to make available on a web dashboard the result of the embedded processing

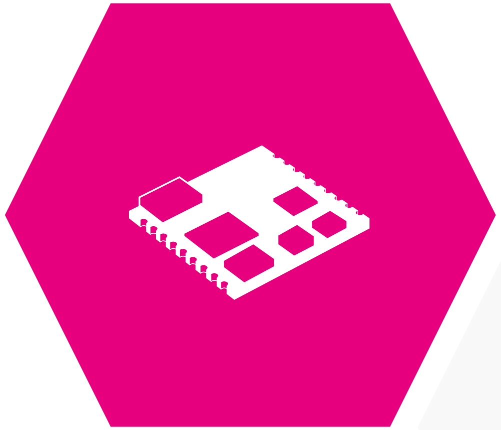
www.st.com/SensorTileBox

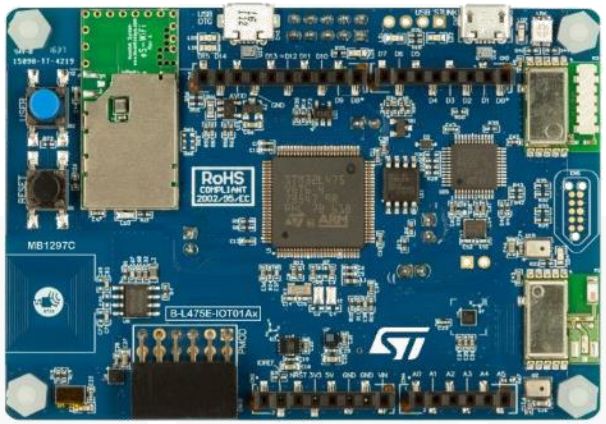
Form factor hardware
AI IoT node for more connectivity

B-L475E-IOT01A

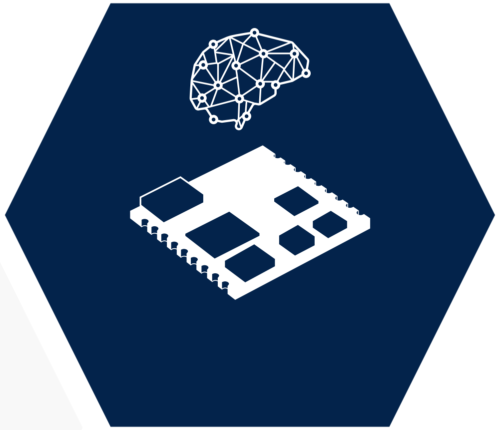
+ Sub-1GHz
Dynamic NFC Tag
Wi-Fi

Capture Data





Inference on STM32L4



More debug capabilities

- Integrated ST-Link/V2.1
- PMOD extension connector
- Arduino Uno extension connectors

www.st.com/IoTnode

Wireless Industrial node to capture data at industrial grade

STWIN

Capture Data

Inference on STM32L4R9

Industrial-grade sensors

- Industrial scale 9-DoF motion sensors including accelerometer, gyrometer and an ultra wide-bandwidth vibrometer with ultra low noise
- Very high frequency audio and ultrasound microphone
- High precision temperature and environmental monitoring
- Micro SD card for standalone data logging
- BLE5.0 connectivity and WiFi expansion board
- USART

www.st.com/stwin

life.augmented

STM32H7 discovery boards with camera

STM32H747I-DISCO
with B-CAMS-OMV

Capture Data

Inference on STM32H747

Computer Vision on microcontroller

- STM32H747 high-performance and DSP with DP-FPU, Arm Cortex-M7 at 480 MHz + Cortex-M4 MCU with 2MB internal Flash, 1MB internal RAM, Chrom-ART Accelerator
- External memory 2x64MB Quad-SPI NOR Flash and 32MB SDRAM
- 4” capacitive touch LCD display module with MIPI® DSI interface
- Camera module adapter board and camera module based on OV5640 5MPx 8b color rolling shutter
- ST-MEMS digital microphones
- Ethernet RJ45 and WiFi / cellular expansion boards

life.augmented

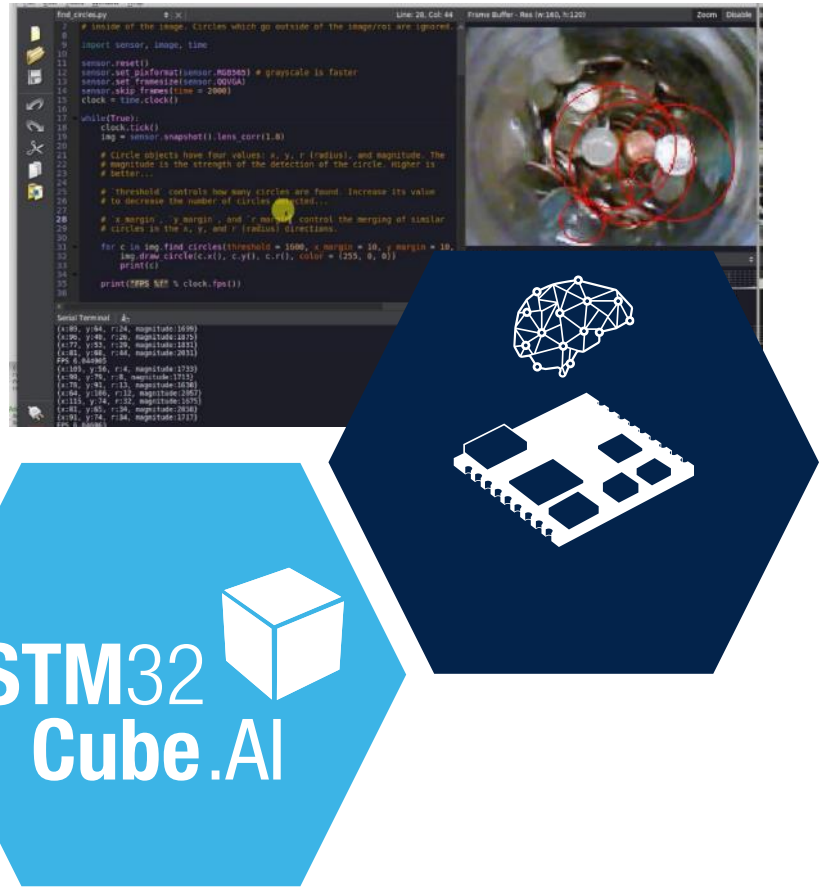
OpenMV integration
Fast machine vision prototyping

OpenMV

OpenMV CAM
Running MicroPython over STM32

Configure Machine Vision in real-time over USB in Python

Run and validate optimized Neural Network

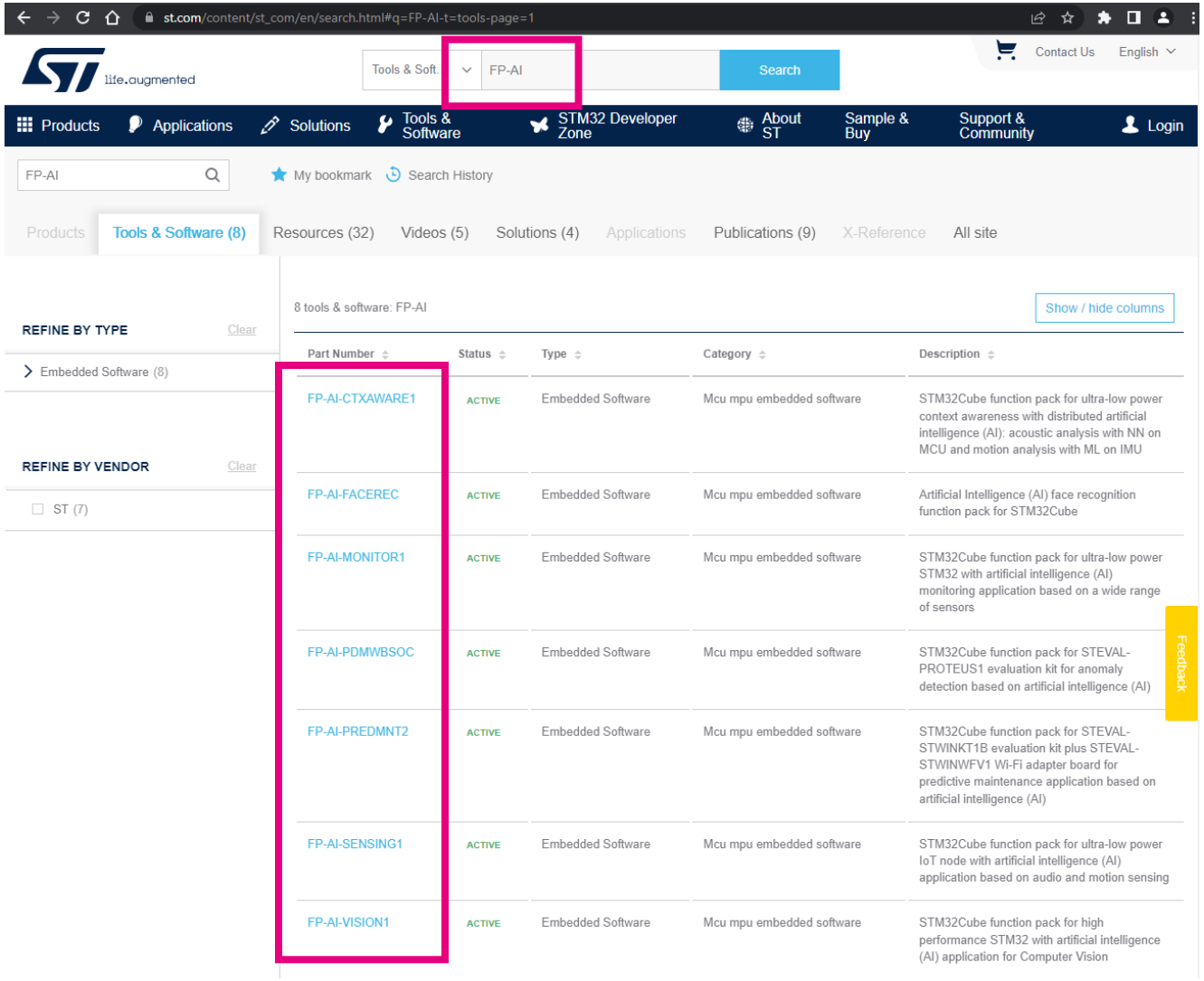


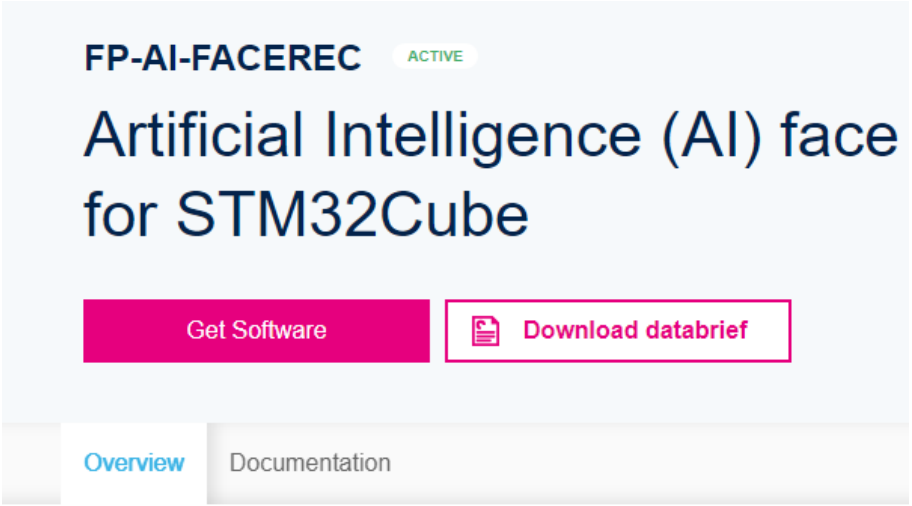
https://github.com/openmv/openmv

AI Application demos

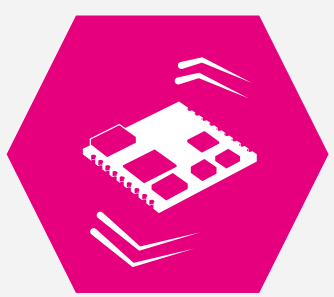
life.augmented

STM32 AI Application demos





Human activity motion recognition (HAR)



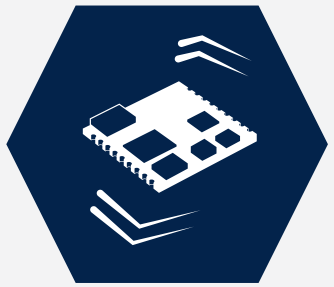
5 classes example

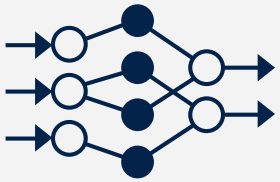
Motion **Data Capture**

Labelling controlled
by smartphone application

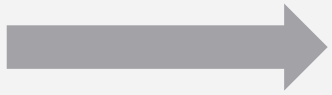
Data stored on the device
SD card for future **learning**

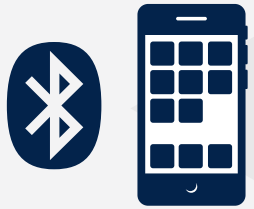
Stationary, walking, running,
biking, driving **labelling**





NN & example
dataset provided





Embedded **motion**
pre-processing

Inferences running
on the microcontroller

Inference result
displayed on mobile app

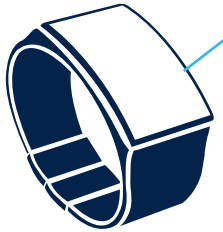
life.augmented

Human Activity Recognition

Classify human activity in wearables

Use cases

- Adapt behavior based on user's activity environment: switch peripherals to low power modes
- Basic heath assistant measuring time spent moving per day



Smartwatch enables BLE when user is driving

Sport watch monitor time spent on sport activities

SensorTile.Box

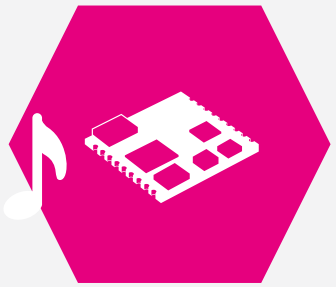
Demo overview

- Activity is displayed in BLE mobile app
- Ultralow-power always-on Human Activity exploits 3-axis accelerometer data
- 5 classes: stationary, walking, running, biking, driving
- Pre/Post-processing: filtering gravity, reference rotation, temporal filter

STM32 Cube.AI Neural Network on STM32L4

life.augmented

Audio scene classification (ASC)



3 classes

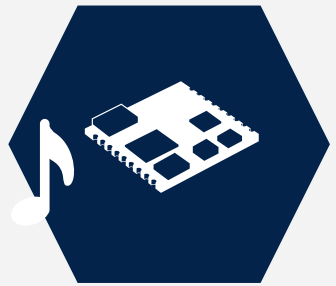
Audio **Data capture**

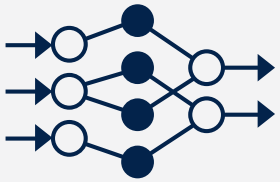
Labelling controlled
by smartphone application

Data stored on the device
SD card for future **learning**

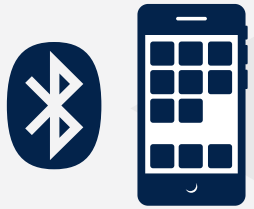
Indoor, Outdoor, In vehicle

labelling





NN & example
dataset provided



Embedded **audio**
pre-processing

Inferences running
on the microcontroller

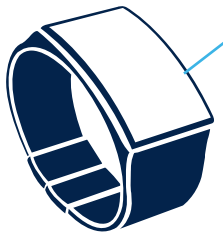
Inference result
displayed on mobile app

Acoustic Scene Classification

Classify acoustic scene

Use cases

- Adapt behavior based on user's environment: switch peripherals to low power modes
- Detect time spent driving per day



Smartwatch turns off WIFI when not indoor

Sport watch turns GPS when outdoor

SensorTile.Box

Demo overview

- Exploits 16kHz microphone data
- 3 classes: indoor, outdoor, in-vehicle
- Pre-processing: LogMel spectrogram on-board

STM32 Cube.AI Neural Network on STM32**L4**

▶ Demo available on FP-AI-SENSING1

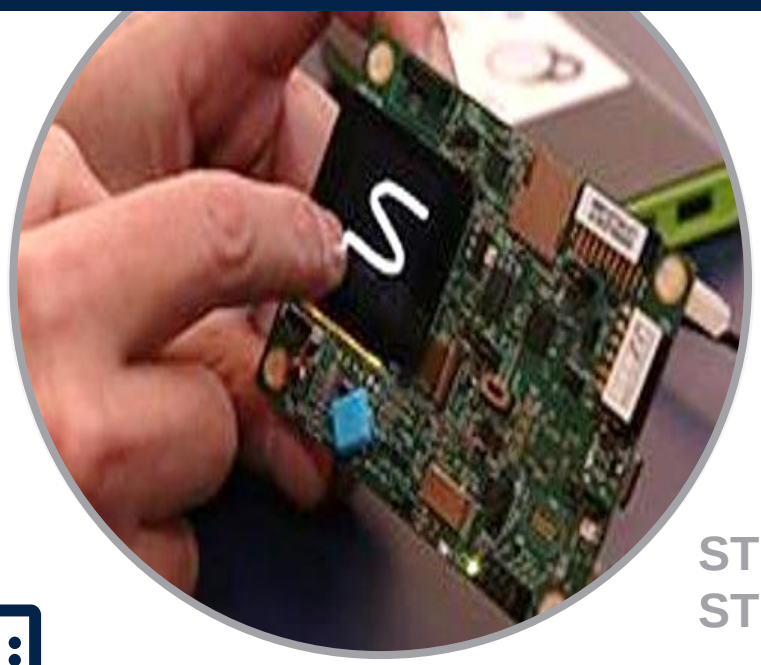
life.augmented

Handwritten character recognition

Character recognition on touchscreen for wearable HMI

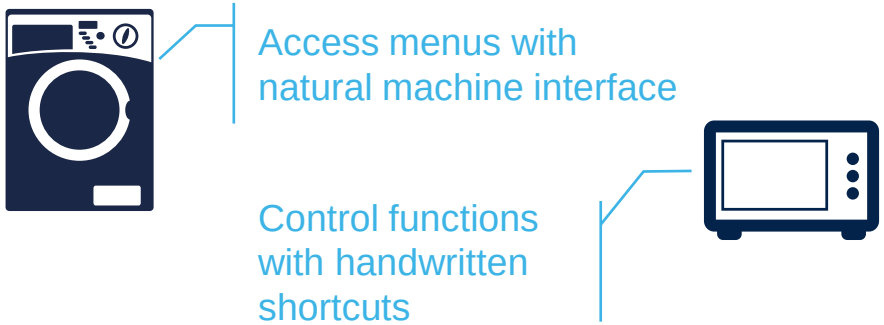
Use cases

- Neural Network for wearables HMI improvement
- Each character or word can be associated to actions



Demo overview

- Handwritten character recognition from touch screen panel
- Screen panel screenshots 32x32 images provided to the Neural network



STM32L5 DK
STM32L496 DK

STM32 Cube.AI Neural Network on STM32**L4** STM32**L5**

life.augmented

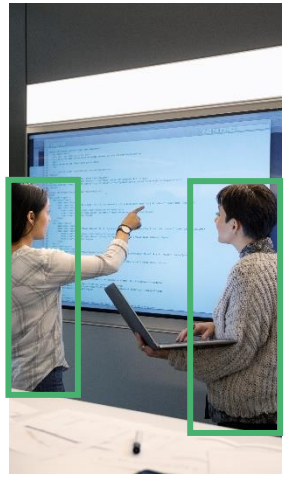
Computer Vision for STM32

Give vision to your STM32 product for new features and add-on services

Food classification

Person presence detection

Face recognition



People counting

FP-AI-VISION1 v1.0	FP-AI-VISION1 v2.0	FP-AI-FACEREC v1.0	FP-AI-VISION1 v3.0

Keras TensorFlowLite ONNX

PyTorch And more

STM32 Cube.AI ⟷ STM32 CubeMX

Optimized NN files

STM32.AI lib

Customer application

STM32.Vision lib

STM32 **run-time**

life.augmented

Food classification

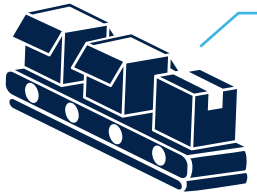
Classify live image stream in real-time

Use cases

- Classify image content to trigger custom actions or content-specific notifications
- Runs in medium FPS allowing real-time actions
- Validated with RGB camera, can be extended to other types

Demo overview

- Camera acquisition to image pre-processing before feeding the NN
- Multiple memory mapping possibilities to optimize / test impact on perf
- Quantize your trained network to optimized inference time and memory usage



Route packages in supply chain based on their standardized appearance

Sort waste and assign to correct containers

STM32H747 Discovery + F4DIS-CAM

STM32 Cube.AI Neural Network on STM32**H7**

life.augmented

Facial expression recognition

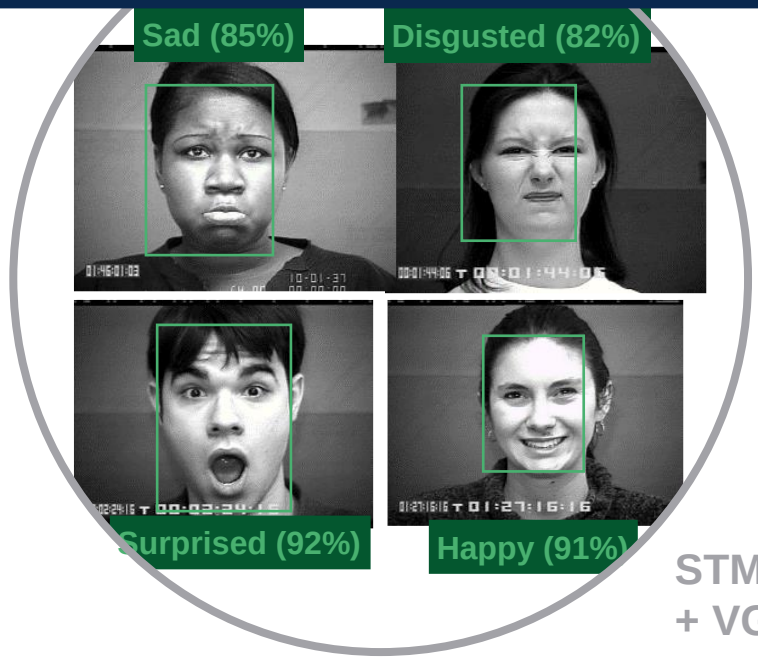
Live face expression recognition from a camera video stream

Use cases

- Build interactive experience in toys
- Optimize advertising experience based on customer emotional response
- Only a low-resolution camera is required

Assert customer interest at point of sales, like vending machines

Improve toys reaction based on user's expression



Demo overview

- Multiple face detection with labeled bounding box displayed in real-time
- Recognize 7 expressions: angry, disgusted, scared, happy, neutral, sad, surprised
- Processing time: 4.8 FPS for 1 face, 3.4 FPS for 2 faces

STM32H747 Discovery
+ VG5661 (or F4DIS-CAM)

STM32 Cube.AI Neural Network on STM32**H7**

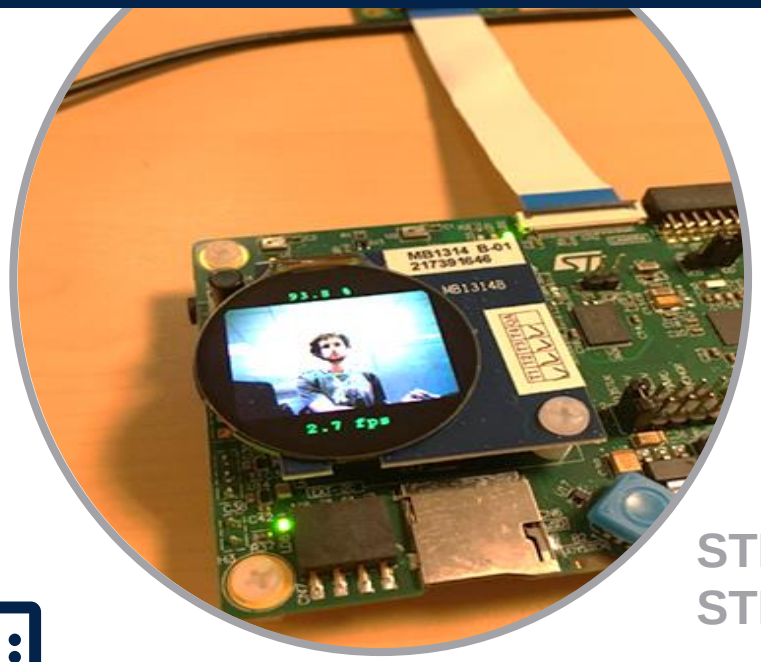
life.augmented

Person presence detection

Detect person presence even without movement

Use cases

- Visual wake word for Smart home or cities security cameras
- Reduce false alarms due to object movement detection



L4R Demo overview

- Person Presence Detection model from tensorflow.org (L4R and H7) and MobileNet v2 (H7 only)
- QVGA 320x240 color image on the LCD
- Can adapt camera flipping depending on which side camera is placed

STM32L4R DK
STM32F4DIS-CAM

Trigger alarm on person detection instead of movement

Visual wake word for more advanced person identification

STM32 Cube.AI Neural Network on STM32**L4R**

life.augmented

Anomaly Detection using vibration

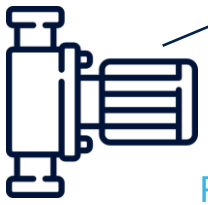
Detect anomaly in industrial equipment using monitoring device

Use cases

- Key component of condition monitoring system for industrial equipment with rotating parts: motors, fans, pumps, bearings, spindles…
- Detect anomalies in complex system based on 3-axis accelerometer sensor
- No training on abnormal state is required, on-chip anomaly detection is done in real-time

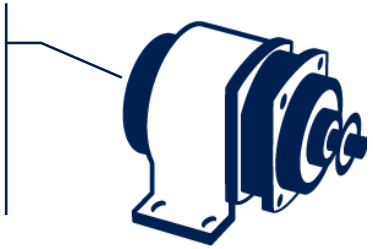
L5R Demo overview

- Function Pack provides logging capabilities, display and local control via USB CLI.
- NanoEdge AI™ Studio by Cartesiam to generate custom machine learning library able to learn and infer on STM32
- Learn normal state on STM32 during onsite initialization phase, then detect anomalies from vibration



Real-time detection of anomalies of industrial pumps

Real-time detection of shocks, misalignment or motor unbalance on industrial appliances



Motor Control Kit G4
STM32L5R Discovery Kit

STM32 Cube.AI + NANOEDGE AI™ by CARTESIAM — Neural Network on STM32**L5R**

life.augmented

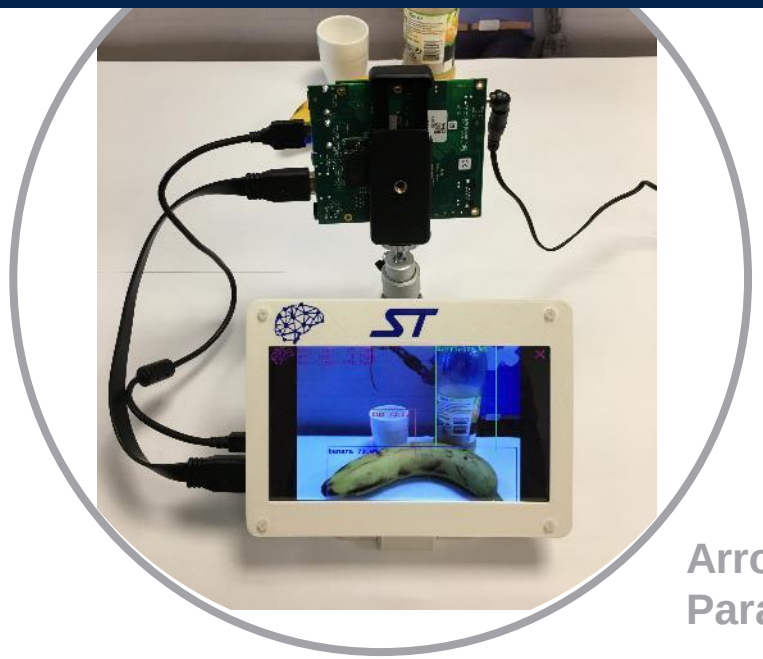
Multiple object detection on STM32MP1 MPU

Advanced object detection among 90 different objects using TensorFlow Lite on STM32MP1

Use cases

- Detect object and its position.
- 90 objects can be detected
- Object detection is done in real time for fast interaction with user
- Requires only a low resolution camera

Identify and locate potential instances of plant disease



Demo overview

- TensorFlowLite integrated via C++ runtime implementation on dual-core A7
- COCO ssd MobileNet v1: 90 classes
- CPU load balanced on the 2 cores
- Processing time: 1.1 FPS

Arrow Avenger96 board
Parallel interface Camera Display

STM32 Cube.AI Neural Network on STM32**MP1**

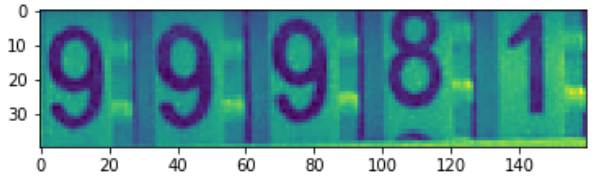
life.augmented

Aftermarket wireless digit reader for metering

Equip meters with aftermarket Wireless & Low power reader

Use case

- Equip meter with ad-on SPI low-cost camera
- Boost ROI avoiding onsite visit implementing long range wireless reader
- Electrical, gaz, water meters supported
- Reading sent over LoRAWan



Demo overview

- QQVGA input @4fps
- Proprietary Neural Network trained on SCUT WMN dataset
  - Accuracy: 95%
  - Inference 84 ms / digit
  - Flash: 16 kB, RAM : 8 kB
- LoRaWAN stack on LoRa SoC
  - Flash: 65KB, RAM: 6KB

STM32WL55JC
Arducam mini 5MP plus

STM32 Cube.AI Neural Network on STM32WL

life.augmented

Why choose STMicroelectronics?
A complete ST AI solution stack

life.augmented

A full development ecosystem
Create your AI application on STM32

FP-AI-VISION1

Computer Vision MCU applications

FP-AI-SENSING1

Audio and Sensing MCU applications

FP-AI-MONITOR1

Time series-based MCU monitoring applications

X-LINUX-AI

AI frameworks on STM32 Microcontrollers

Various **application examples** to get started rapidly

Data processing libraries to speed up development

Step-by-step guides to extend examples or benchmark ML performance

STM32 **Community** with dedicated Neural Networks topic and **AI expert partners**

Trainings, hands on, MOOCs and partners **videos**

life.augmented

Whatever your company's profile, you will find an AI solution suited to your needs

COMPANY'S PROFILE	USE CASES		
	Anomaly detection	Classification	Deep Learning
Embedded developers ▪ No dataset available ▪ No dedicated AI Team	NANOEDGE AI STUDIO	NANOEDGE AI STUDIO	Engineering Services STM32 Cube.AI
Team with AI expertise ▪ Dataset available ▪ AI Team		NANOEDGE AI STUDIO STM32 Cube.AI	STM32 Cube.AI

life.augmented

Don't go alone

A network of companies is there to support you

Trust our **authorized partners** to ensure the success of your project. Learn more at st.com/stm32ai

Would you like to discuss a co-development partnership for ML/AI projects? Contact us at edge.ai@st.com

life.augmented

Releasing your creativity

- /STM32
- @ST_World
- community.st.com
- www.st.com/STM32ai
- wiki.st.com/stm32
- github.com/STMicroelectronics
- Videos
- STM32Cube.AI blog articles

life.augmented

Our technology starts with You

Find out more at www.st.com

life.augmented