



life.augmented

STM32 Enhanced AI Solutions and Ecosystem

STMicroelectronics

문현수 과장

Introduction to artificial intelligence at the edge

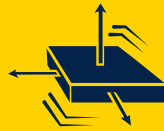


Signals turning into data

Embedded applications will collect more data in the future



Growing demand for data-driven insights



Increasing use of sensors



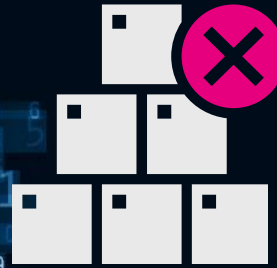
Proliferation of IoT devices



AI offers the best approach to process this growing amount of data



Algorithms and predefined models to analyze data and make predictions or decisions



Traditional approaches have limitations:

- when dealing with **large datasets**
- when the **phenomena are too complex**



Machine learning algorithms to automatically learn patterns and relationships from data



AI-based data processing offers a more flexible and powerful approach to analyzing and making decisions from large data collections

The rise of Edge AI



Ultra-low latency
Real-time applications

01 **Reduced data transmission**
10 Generate meaningful information



Enhanced privacy and security
No data sharing in the cloud



Power efficiency
Low-data / Low-power



Improved accuracy
analyze data from a wide range
of sensors and sources

Edge AI benefits many application domains:

Industrial maintenance

Condition monitoring
Predictive maintenance



Control systems

From home heating systems
to industrial machines



Internet of Things (IoT)

Smart cities, smart buildings,
connected homes, and
industrial automation

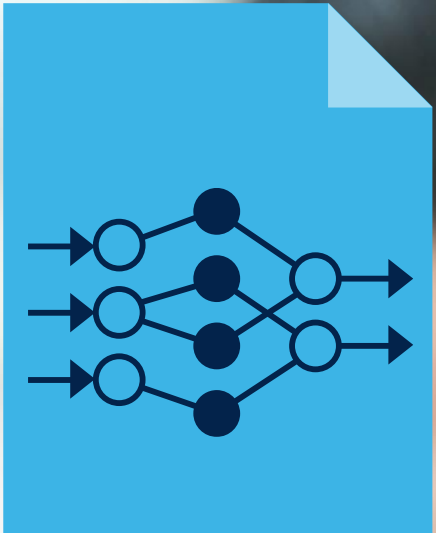




STM32 is a key enabler

**Democratizing edge AI
with STM32 platforms**

STM32 Edge AI



Power consumption

Memory footprint

Inference time

Edge AI expertise

eSW development

Data

Edge AI development workflow

ST software offering

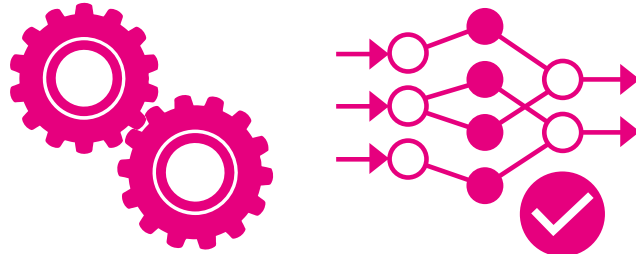
1 Data Preparation



Data acquisition

Data processing

2 Data Science



Model selection and training

Model validation

3 Model Implementation



Model library creation

Model inference

NANOEDGE AI STUDIO 

Automated edge AI software

STM32 
Cube.AI

Edge AI toolkit



All STM32 MCUs



STM32 product offering simplifies developers' approach to edge AI

- 3 products for embedded developers and data scientists

AD + LOD
SENSING

NANOEDGE AI STUDIO

User-friendly Auto-ML tool for STM32 MCUs

SENSING
AUDIO
VISION

STM32 Cube.AI

AI model optimizer and code generator for STM32 MCUs

SENSING
AUDIO
VISION

X-LINUX-AI

A complete AI framework for STM32 MPUs

- Covering a broad variety of applications

ANOMALY DETEC. + L.O.D.

- Anomaly detection
- Predictive maintenance
- Learning on device

SENSING

- Sensor analysis
- Activity recognition (motion sensors)
- Stress analysis or attention analysis

AUDIO

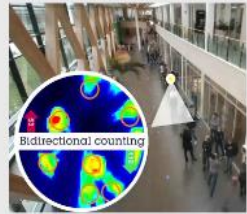
- Audio (key word, scene detection)
- Speech (sentences) recognition
- Speech synthesis

VISION

HORSE + DOG + CAT

- Multiple object detection
- Face/object analysis (face detection)

A proven technology already widely adopted



Schneider
Electric

STM32
Cube.AI

SMART OFFICE | CUSTOMER

People flow counting Sensor with Schneider Electric

An innovative approach to measure people flows using an in-house thermal sensor.



STM32
Cube.AI

INDUSTRIAL | DEMO

Aftermarket wireless digit reader

Equip meters with aftermarket wireless & low-power readers.



STM32
Cube.AI

SMART CITY | DEMO

Acoustic scene classification

Identify different environments (indoor, outdoor, in-car) using a simple microphone.



STM32
Cube.AI

INDUSTRIAL | DEMO

People presence detection (visual wake word)

Human detection on high-performance MCU.



STM32
Cube.AI

SMART HOME | DEMO

Floor type detection for vacuum cleaners

Advanced solution for material recognition of floor type (hard or soft) enabled by AI technology.



STM32
Cube.AI

INDUSTRIAL | DEMO

Fan anomaly classification based on ultrasound analysis

Neural Network classification based on a high-frequency analog microphone pipeline.



STM32
Cube.AI

SMART BUILDING | DEMO

Hand posture recognition without camera module

Hand posture recognition running on STM32F401 based on ST multizone Time-of-Flight ranging sensor.



STM32
Cube.AI

WEARABLES | DEMO

Human Activity Recognition

Easily identify 5 different activities with a 3D accelerometer.

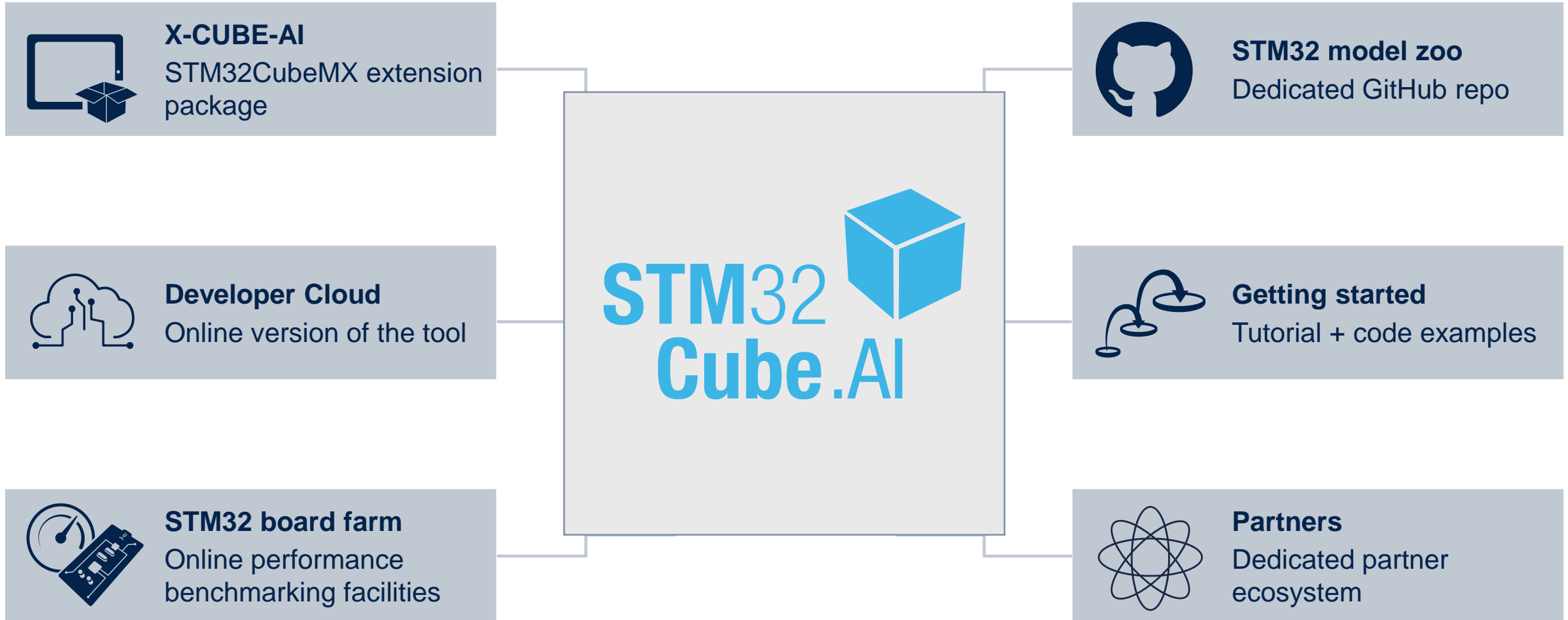
STM32Cube.AI

Edge AI optimization tool for STM32

The screenshot displays the STM32Cube.AI Developer Cloud interface. At the top, a progress bar shows five steps: Optimize, Configure, Benchmark, Results, and Generate. The 'Optimize' step is currently active. Below the progress bar, the 'Model currently selected' dropdown shows 'MOBILENET_18_025.H5'. A table below this shows the current configuration: INPUT (float 1x224x224x3), OUTPUT (float 1x1x1x1), MODEL TYPE (float), and MACC (1267194). A 'Select another model' button is located below the table. The 'Select your model optimization options' section contains four radio buttons: 'Balance between RAM size and inference time (-optimization balanced)' (selected), 'Optimize for RAM size (-optimization ram)', 'Optimize for inference time (-optimization time)', 'Use activation buffer for input buffer (-allocate-inputs)', and 'Use activation buffer for output buffer (-allocate-outputs)'. An 'Optimize' button is at the bottom right of this section. The 'History of optimization results' section shows a table with columns: Date, Optimization, Allocate Inputs, Allocate Outputs, MACC, Flash Size, and RAM Size. The table contains one row of data for a 'balanced' optimization on 1/19/23 at 6:22 PM.

Date	Optimization	Allocate Inputs	Allocate Outputs	MACC	Flash Size	RAM Size
1/19/23, 6:22 PM	balanced	true	true	1267194	541548	617104

STM32Cube.AI solutions









STM32Cube.AI core technology

 **STM32 model zoo**





Bring your own model (BYOM)

 via 	 via 	 MATLAB via 
---	---	---




Optimize and validate your NN model

STM32 Cube.AI




STM32Cube.AI for desktop

 STM32Cube ecosystem	 Command Line Interface
--	---

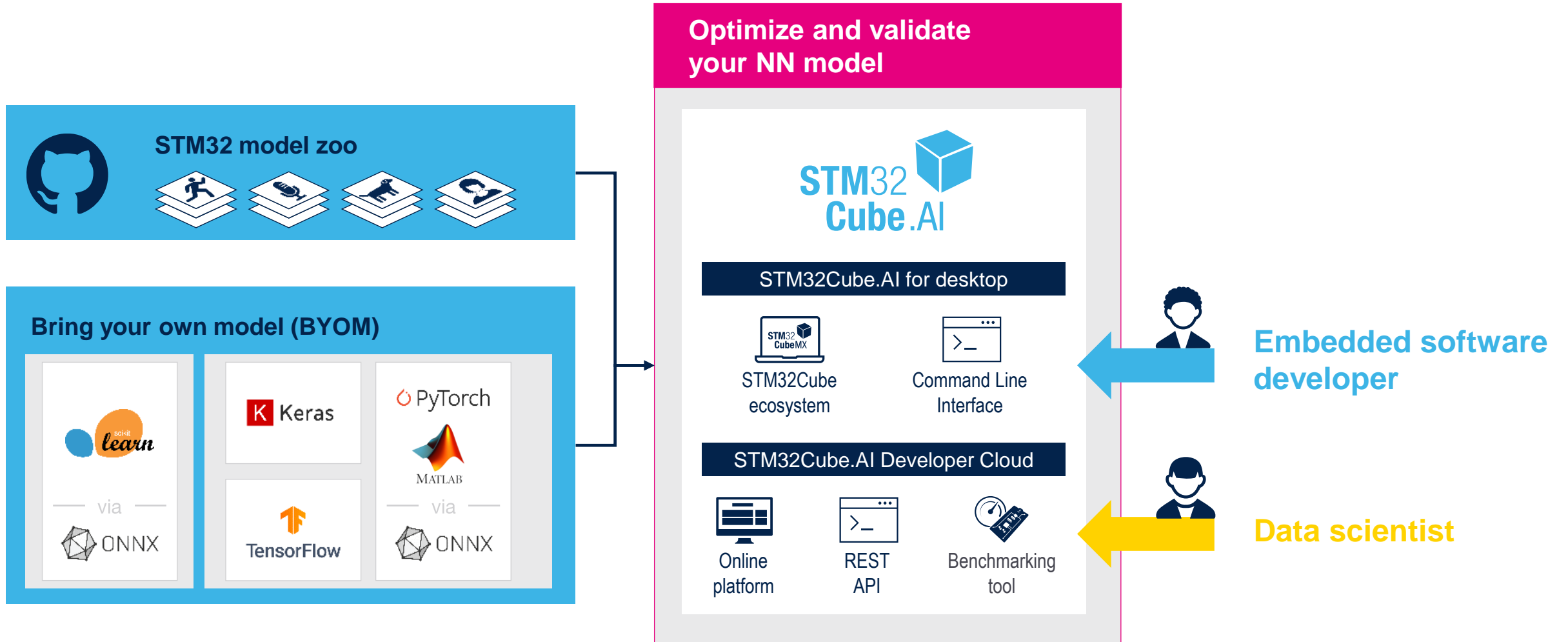
STM32Cube.AI Developer Cloud

 Online platform	 REST API	 Benchmarking tool
--	---	--

STM32 Cube.AI Core technology

-  Evaluate
-  Optimize
-  Finetune

Two versions of the same tool depending on your profile



STM32 benchmarking tool

The unique possibility to evaluate the performance of models remotely, on real STM32 boards



Get the real inference time from optimized models running on STM32



Benchmark models on a large variety of STM32 boards

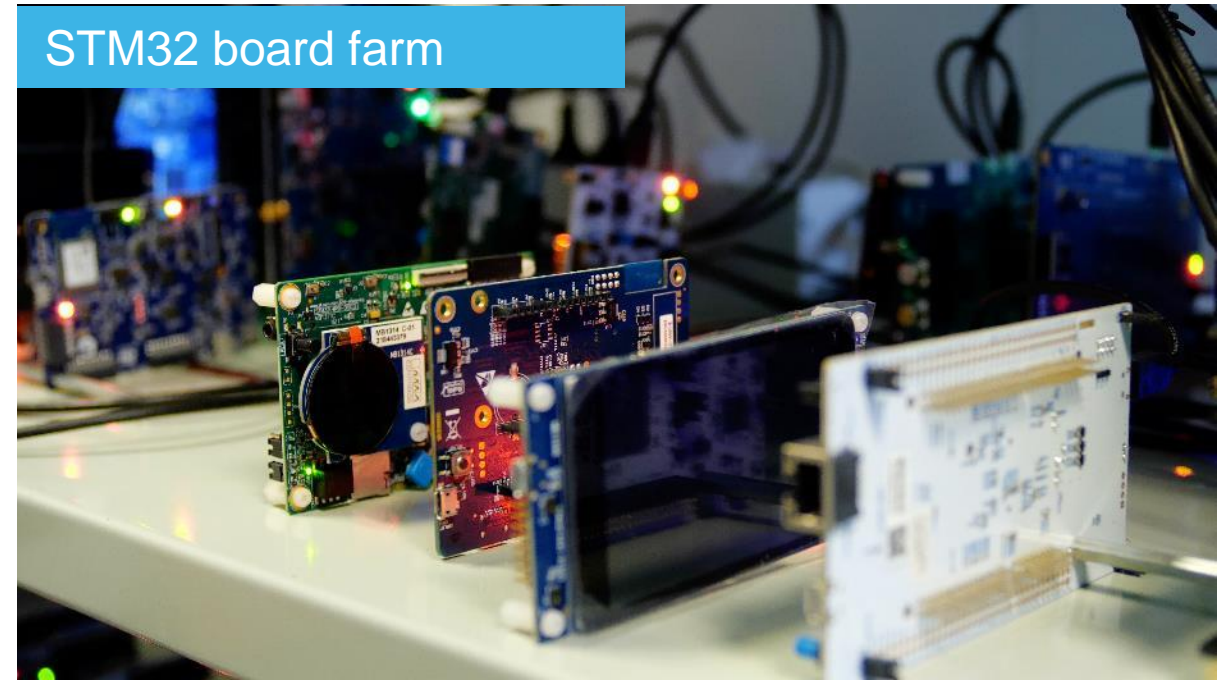
Find the most appropriate board for your application



Get access to the most recent devices

A board farm is constantly updated with the latest available boards

STM32 board farm



STM32 model zoo

A collection of application-oriented models optimized for STM32

Human activity



Motion Sensing

Image classification



Computer vision

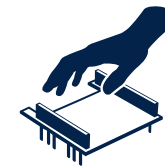


Hosted on GitHub



Model training scripts

- Scripts to generate and validate



Getting started application packages

- Automatically generated from the trained models
- Easy to deploy for end-to-end evaluation

Audio event detection



Audio classification

Object detection



Computer vision



STM32Cube.AI can run on all STM32 series

Five product categories



Wireless
MCU

Short- and long-range connectivity



Ultra-low-power
MCU

32-bit general-purpose microcontrollers: from 75 to 3,224 CoreMark score



Mainstream
MCU



High-performance
MCU



Embedded
MPU

32- and 64-bit microprocessors



Enabling edge AI solutions

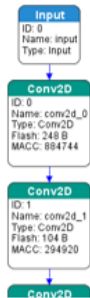


Scalable security

The 3 pillars of STM32Cube.AI

Graph optimizer

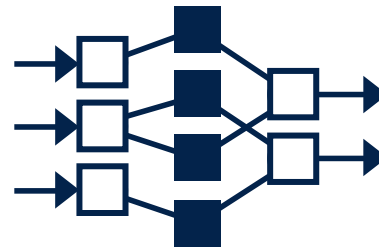
Automatically improve performance through graph simplifications & optimizations that benefit STM32 target HW architectures



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

Quantized model support

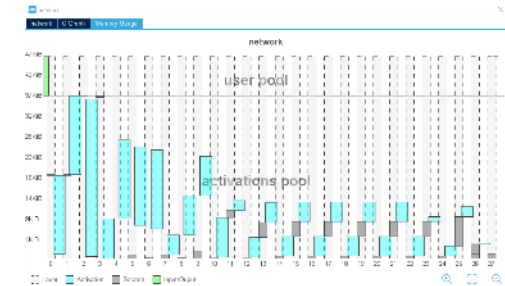
Import your quantized ANN to be compatible with STM32 embedded architectures while keeping their performance



- From FP32 to Int8 or mixed-precision
- Minimum loss of accuracy
- Code validation on target
 - Latency
 - Accuracy
 - Memory footprint

Memory optimizer

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

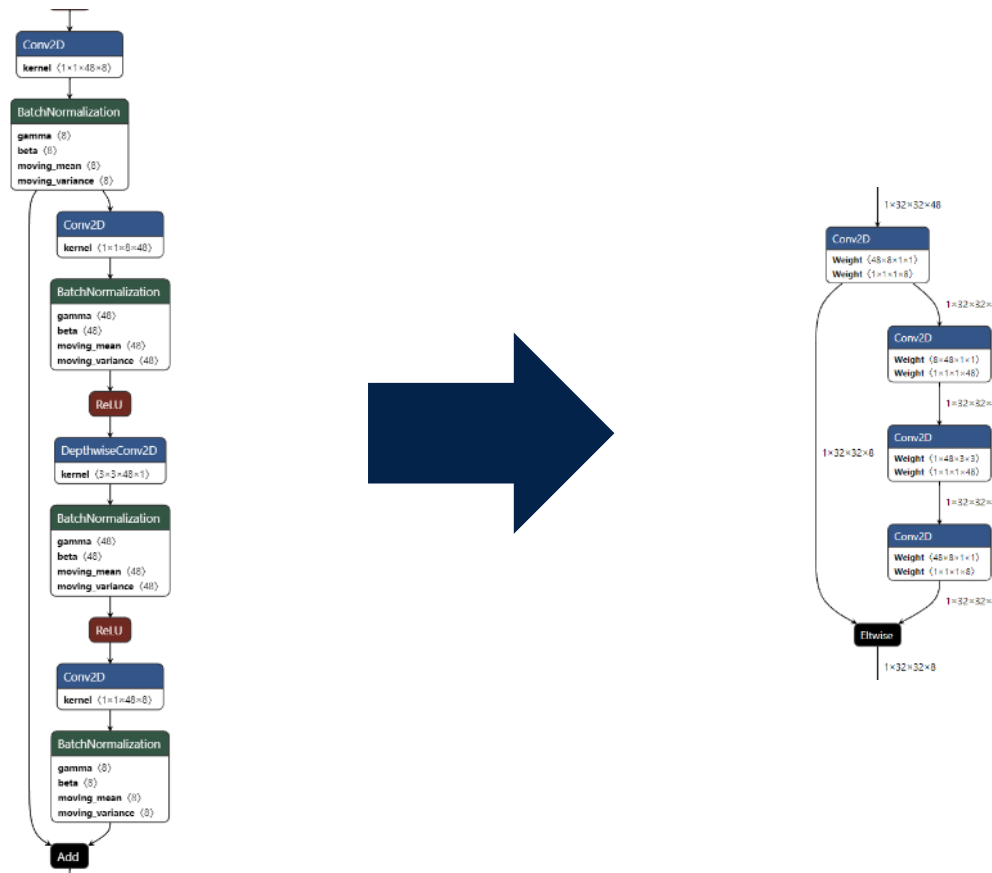


- Memory allocation
- Internal/external memory repartition
- Model-only update option

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

Graph optimizer

Squeeze your graph to fit into an MCU!



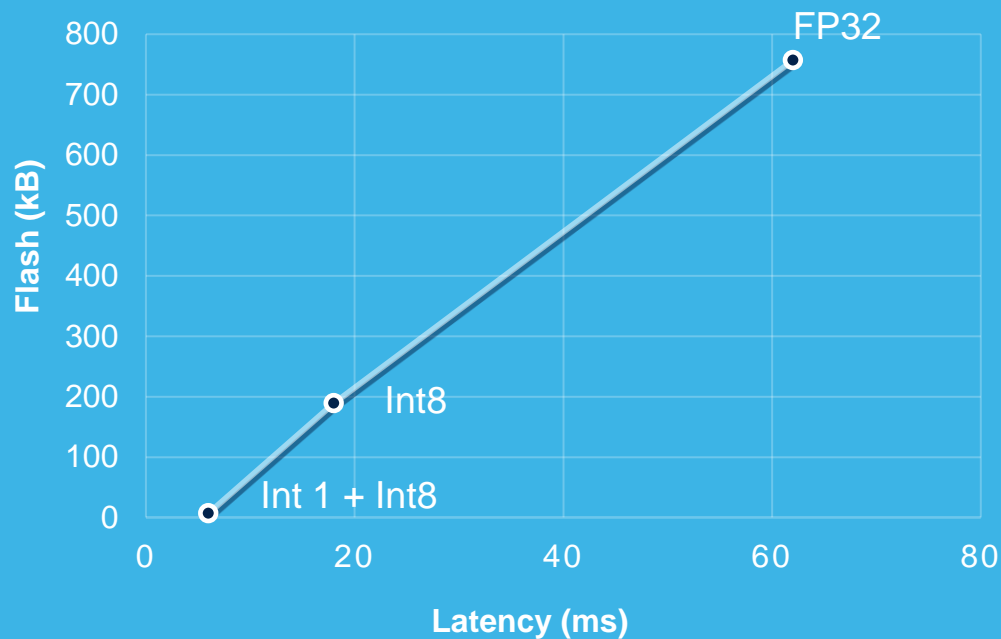
Fully automated process in the STM32Cube.AI workflow

- Your original graph is optimized at the very early stage for optimal integration into the STM32 MCU/MPU
- Loss-less conversion

Quantized model support

Simply use quantized networks to reduce memory footprint and inference time

LATENCY & MEMORY COMPARISON FOR QUANTIZED MODELS



STM32Cube.AI supports quantized neural network models with **all parameter formats**:

- FP32
- Int8
- Mixed binary Int1 to Int8 (Qkeras*, Larq.dev*)

*Please contact edge.ai@st.com to request the relevant version of STM32Cube.AI



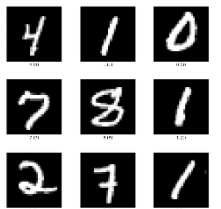
HW Target: NUCLEO-STM32H743ZI2

Model: Low complexity handwritten digit reading

Freq: 480 MHz

Accuracy: >97% for all quantized models

Tested database: MNIST dataset



MNIST dataset

Memory optimizer

Optimize performance easily with the memory allocation tool



Model RAM consumption per layer

- Easily identify the most critical layers

Model memory allocation

- Set your external memory
- Map in non-contiguous internal flash section
- Partition internal vs external flash memories

Re-use model input buffer to store activation data*

- Minimize RAM requirements

Relocatable network

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

Use external flash Memory: Custom

Split weights between internal and external flash using a linker script

Start Address: 0x00000000 Size (Mbytes)

Tensor	Size	Internal 440KB	External 0KB
conv1_weights	864	<input checked="" type="checkbox"/>	<input type="checkbox"/>
conv1_bias	32	<input checked="" type="checkbox"/>	<input type="checkbox"/>
conv_dw_1_weights	288	<input checked="" type="checkbox"/>	<input type="checkbox"/>
conv_dw_1_bias	32	<input checked="" type="checkbox"/>	<input type="checkbox"/>
conv_pw_1_weights	512	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Use external RAM Memory: Custom

Start Address: 0x00000000

Use activation buffer

Start Address: 0x00000000 Act. size (by... 752712

Copy weight to RAM

Start Address: Weight size: 451496

Use activation buffer for input buffer (--allocate-inputs) Force classifier validation output (--classifier)

Use activation buffer for the output buffer (--allocate-outputs)

Split weights during code generation (--split-weights)

Generate relocatable network (--relocatable)

Report's output directory

C:\Users\richard\stm32cubemx Browse...

Enable custom layer support

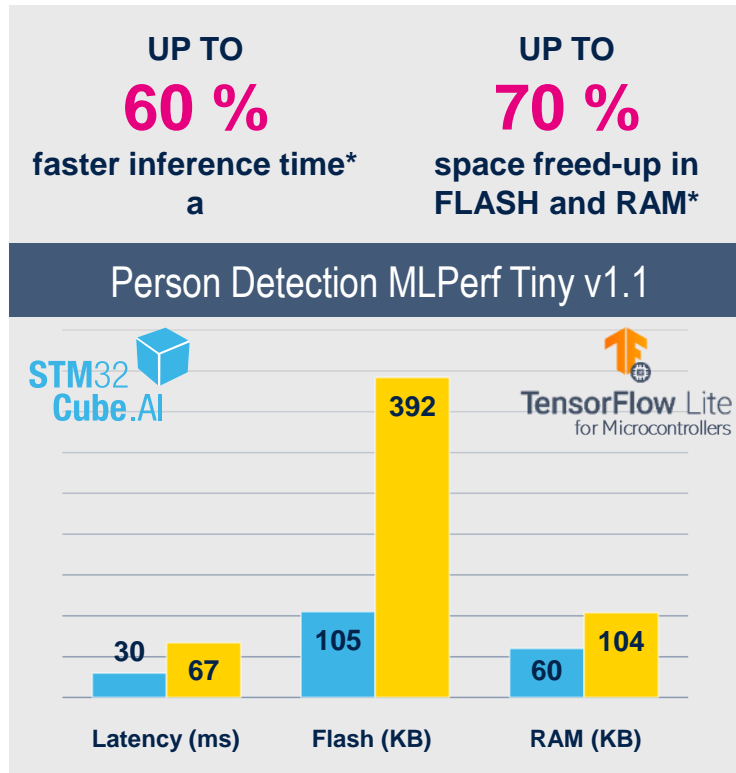
Custom Layer JSON File: Browse...

OK Cancel

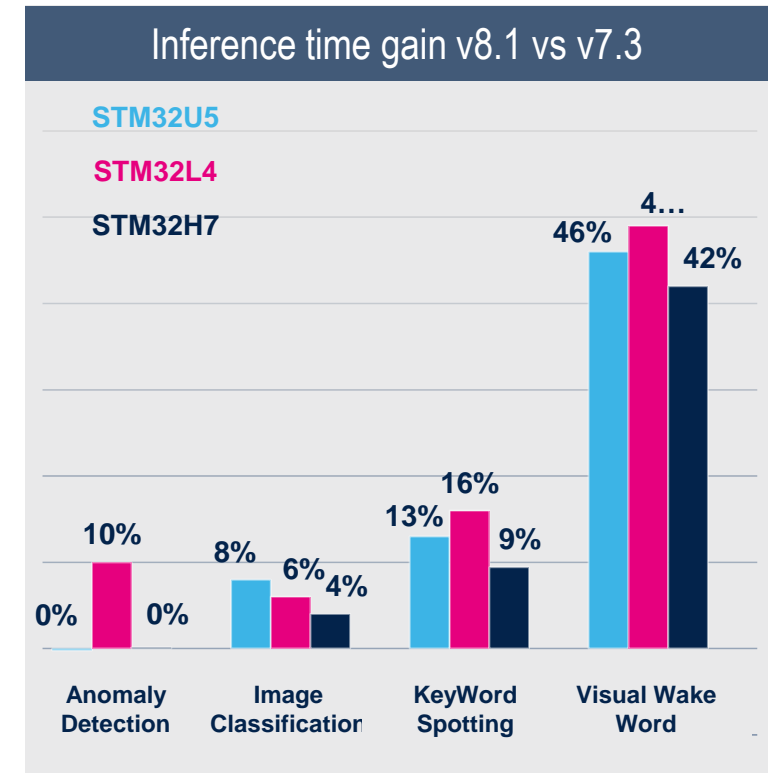
* Requires input and activation buffers in same memory

Pushing STM32Cube.AI performance further

- **Improved performance** in NN models optimization and generation
- **Improved support** of ONNX features and operators
- **Implemented** unsupported layers / operators



* versus TensorFlow Lite for microcontroller using STM32H7A3



We provide everything to kick off your project

Design documentation



Getting started

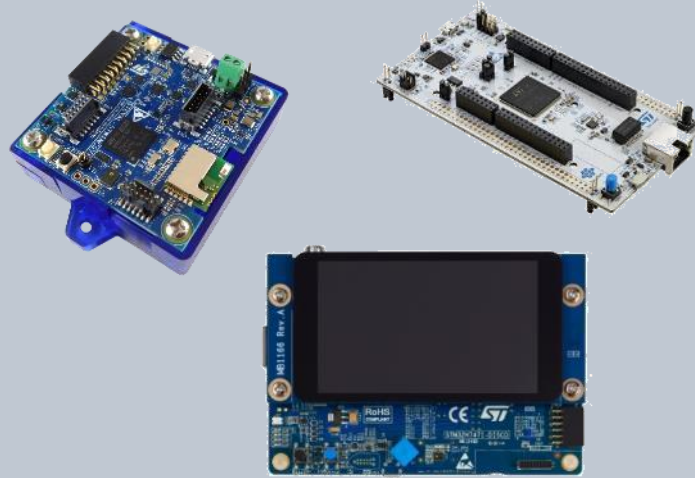
We guide you step-by-step to start with the STM32 ecosystem

Development zone

Get started on application development and project sharing

- **Wiki by ST** is a great forum to learn and start deploying edge AI on STM32!
- Videos of application examples
- Massive open online courses (MOOCs)

Hardware and software tools



- Evaluation platforms for STM32 MCUs and MPUs
- Additional sensor boards
- Full software suite

Support and updates



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

Introducing STM32Cube.AI Developer Cloud



STM32Cube.AI

The original desktop front end AI optimizer for STM32



STM32Cube.AI Dev Cloud

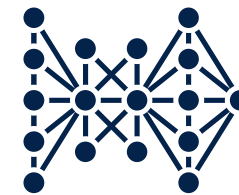
The brand-new online AI services front end for STM32



X-CUBE-AI
for STM32Cube.MX



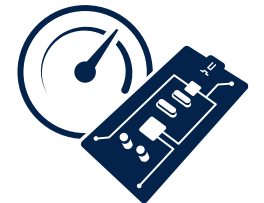
X-CUBE-AI
Command Line Interface



ST Model Zoo



Web GUI
+ REST API

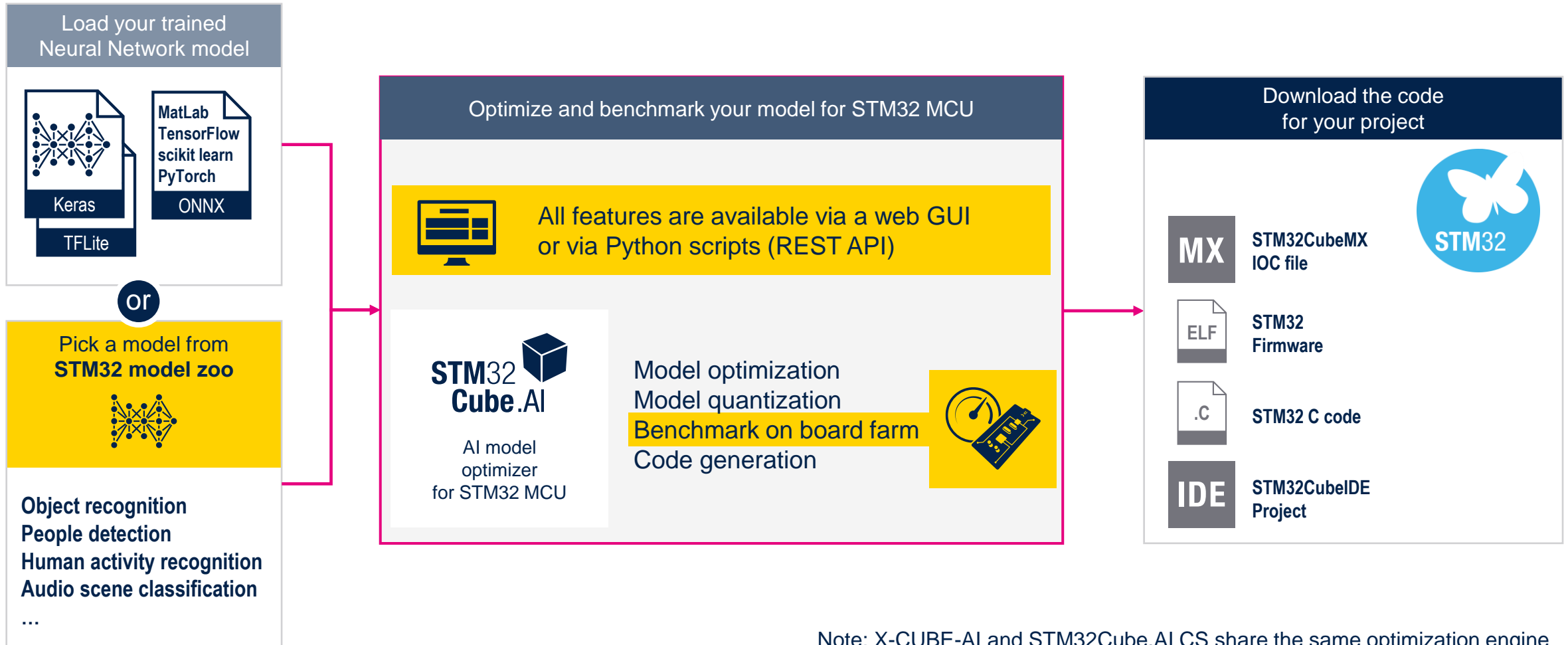


Board farm



Core engine technology

Seamlessly integrate AI in your STM32 projects



STM32 model zoo

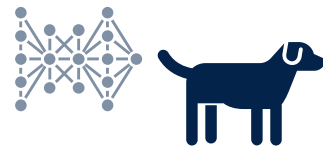
A collection of application-oriented models optimized for STM32

Human activity



Motion Sensing

Image classification



Computer vision

Audio event detection



Audio classification

Object detection



Computer vision



Hosted on Github



Model training scripts

- Scripts to generate and validate

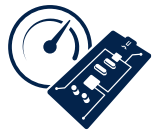


Getting started application packages

- Automatically generated from the trained models
- Easy to deploy for end-to-end evaluation

STM32Cube.AI Developer Cloud Board Farm

The unique possibility to evaluate the performance of models remotely, on real STM32 boards



Get the actual inference time for optimized models



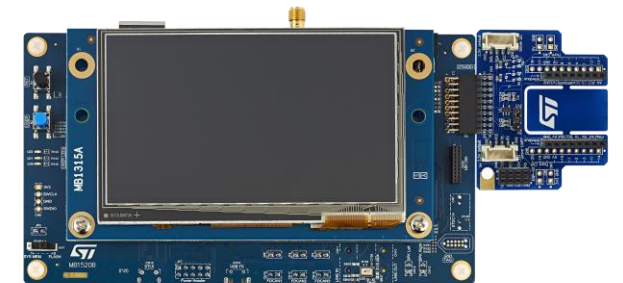
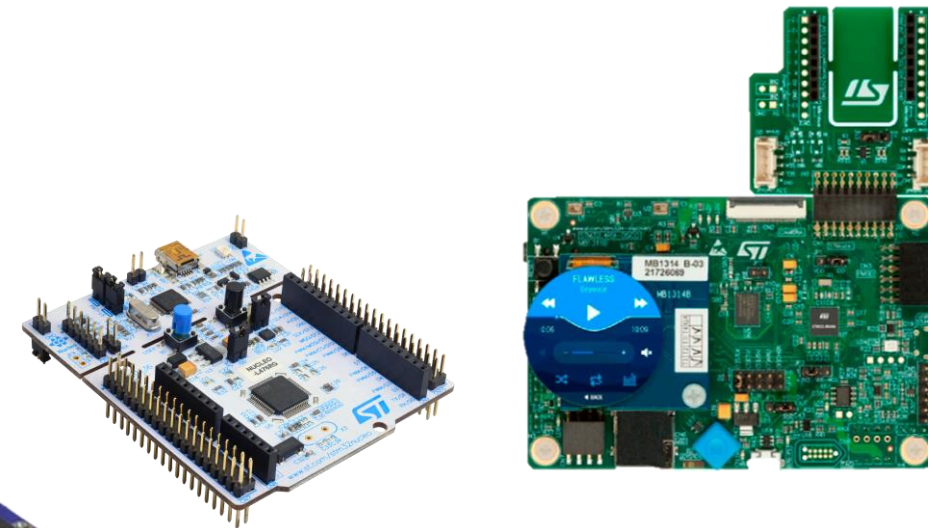
Benchmark models on a large variety of STM32 boards

- Find the most appropriate board for your application



Get access to the most recent devices

- The board farm is constantly updated with the latest available boards



STM32Cube.AI Developer Cloud benefits

STM32Cube.AI Developer Cloud



Save Time

- Immediately start by selecting model from **STM32 model zoo**
- No code/Low code solution with no installation required



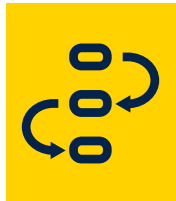
Make the best choice on hardware selection

- Benchmarking service via the **ST board farm**
- Save workload and money



Maximize AI performance on STM32

- Benefit from the STM32Cube.AI performance



Ease integration in your ML workflow

- Use REST API and Python scripts to optimize your ML workflow

NanoEdge AI Studio

Edge AI optimization tool for STM32

The screenshot displays the NanoEdge AI Studio web interface. At the top, a navigation bar includes the STM32Cube AI Developer Cloud logo and a progress indicator with five steps: Optimize, Quantize, Benchmark, Results, and Generate. The main content area shows the 'Optimize' step selected. A dropdown menu indicates the 'Model currently selected' is 'MOBILENET_18_025.H5'. Below this, a table lists model metrics: INPUT (float h(22x22x3)), OUTPUT (float h(1x1x1)), MODEL TYPE (float), and MACC (1267194). A 'Select another model' button is present. The 'Select your model optimization options' section contains three radio buttons: 'Balance between RAM size and inference time (-optimization balanced)', 'Optimize for RAM size (-optimization ram)', and 'Optimize for inference time (-optimization time)'. Two checkboxes are checked: 'Use activation buffer for input buffer (-allocate-inputs)' and 'Use activation buffer for output buffer (-allocate-outputs)'. An 'Optimize' button is at the bottom right of this section. The 'History of optimization results' section includes a table with columns for Date, Optimization, Results, RAM size, MACC, Flash size, and RAM size. The table shows a single entry for '1/19/23, 6:22 PM, Livel, Default' with a 'balanced' optimization, 'true' results, 'true' RAM size, '1267194' MACC, '541548' Flash size, and '617104' RAM size. A 'Show Terminal' button is also visible.

Date	Optimization	Results	RAM size	MACC	Flash size	RAM size
1/19/23, 6:22 PM, Livel, Default	balanced	true	true	1267194	541548	617104

Simplify your AI development workflow

NanoEdge AI Studio, an automated ML design solution

NANOEDGE AI
STUDIO 



The **best combination** for given data:
AI model, hyperparameters and preprocessing

On-device learning capability to fine-tune
deployed solution without retraining

No need to create/train AI models in AI
frameworks

Boost your productivity



Take advantage from
state of the art of ML

AI technological breakthrough
delivered in regular updates








Speed up your
development time

Benchmark thousands of ML
combination in minutes

Generate **ultra optimized ML
library** for every STM32

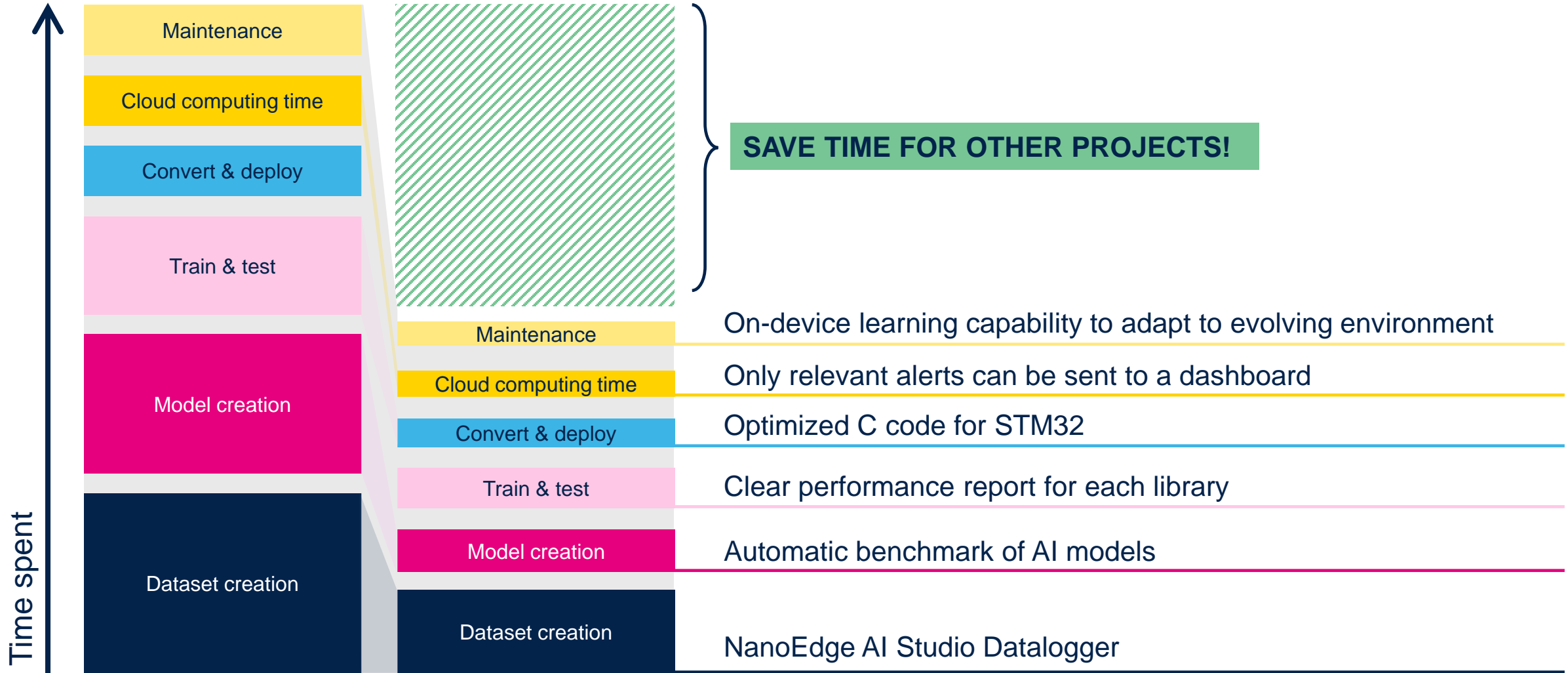
From 0.2 KB to 10 KB
of memory footprint

Create solutions for any STM32 MCU

	Overview	Commercial Part N° ↑	Favorite
<input type="checkbox"/>		ISM330ISN	★
<input type="checkbox"/>		STM32C0	★
<input type="checkbox"/>		STM32F0	★
<input type="checkbox"/>		STM32F1	★
<input type="checkbox"/>		STM32F2	★
<input type="checkbox"/>		STM32F3	★
<input type="checkbox"/>		STM32F4	★

The compact size of the library and its optimized code allows for **seamless integration with any STM32** microcontrollers.

AI solutions development flow enhanced with NanoEdge AI Studio



State-of-the-art machine learning for smarter products



AD
Anomaly
Detection

1C
1-Class
Classification

nC
n-Class
Classification

E
Extrapolation

“

I want to anticipate product failures

“

I need to detect any outliers

“

I want to identify the activity, the environment, the usage

“

I need to predict future states

An integrated workflow

Create your data set
and setup your project

Analyse, choose and validate the best
ML processing

Generate optimized
library for STM32

Train your model directly
on the device



Project parameters
max ram, max flash,
microcontroller ref

01010101
01010101
01010101
01010101

Dataset
Collection of data
from sensors

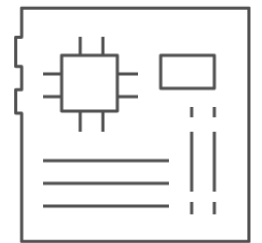


**NANOEDGE AI
STUDIO**

NanoEdge™
AI Studio



**Precompiled
library (.a) to link to
your main code**



on-device
learning

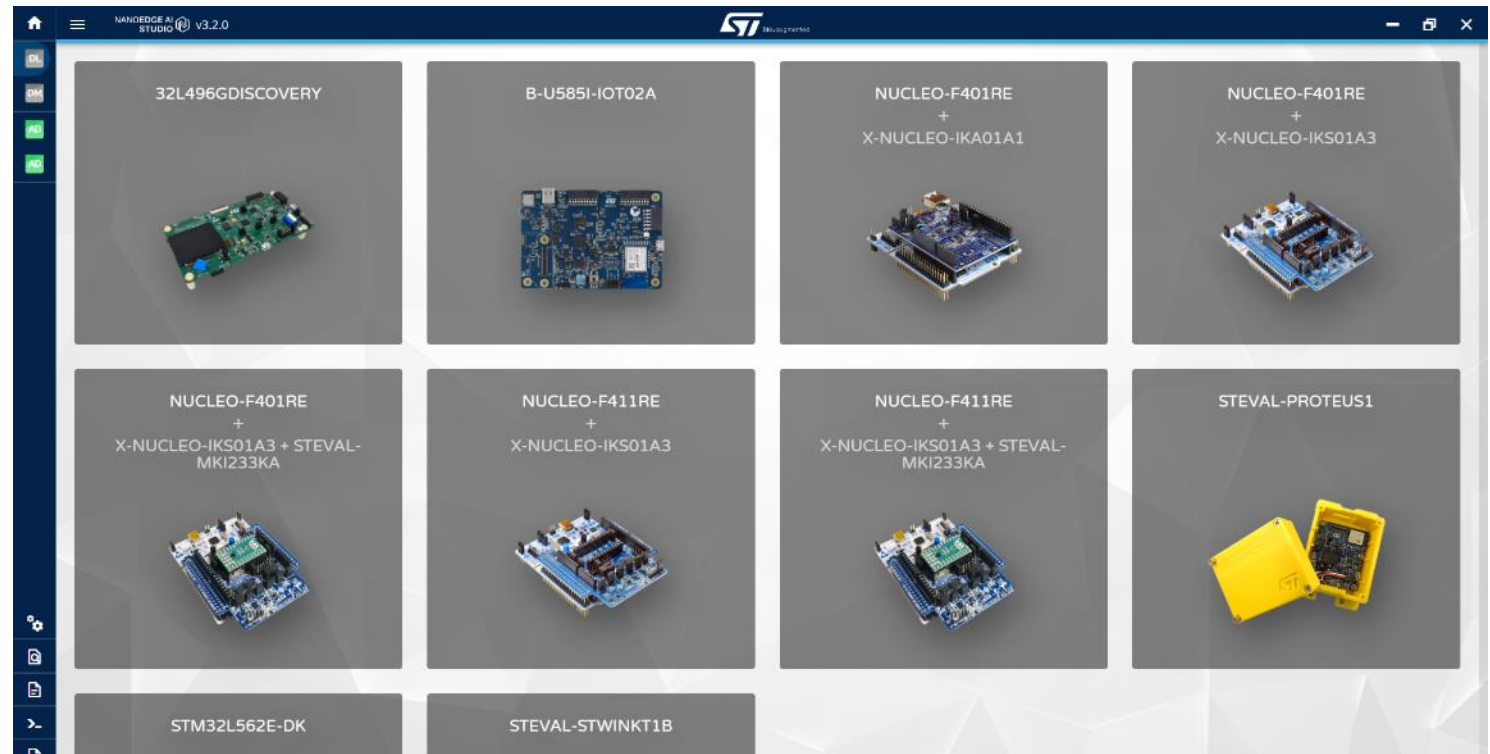
All the tools needed for data collection



DL
Data Logger

“

I want the tools to easily create my dataset”



All the tools needed for data preparation



DM
Data
Manipulation

“

I want to quickly and efficiently clean my data”

The screenshot shows the NANOEDGE AI STUDIO v3.2.0 interface. At the top, there are three tabs: File, Action, and Result, all with checkmarks. The main area is divided into three panels:

- File Panel:** Shows a file named "Log_ISPU_Spacer_2022-06-13-163838.log". It has a toggle for "Ignore first header line" and displays "182 Lines 768 Columns". The "Delimiter" is set to "Space". A "File preview" table is shown below.
- Action Panel:** Shows an "Extract lines" action. A slider is set to 182, and the text "Extract 182 lines" is displayed. A "RUN" button is present. Below the button, a small table shows a preview of the extracted data.
- Result Panel:** Shows the result of the action: "182 Lines 768 Columns". It includes "RUN NEW ACTION" and "SAVE AS" buttons.

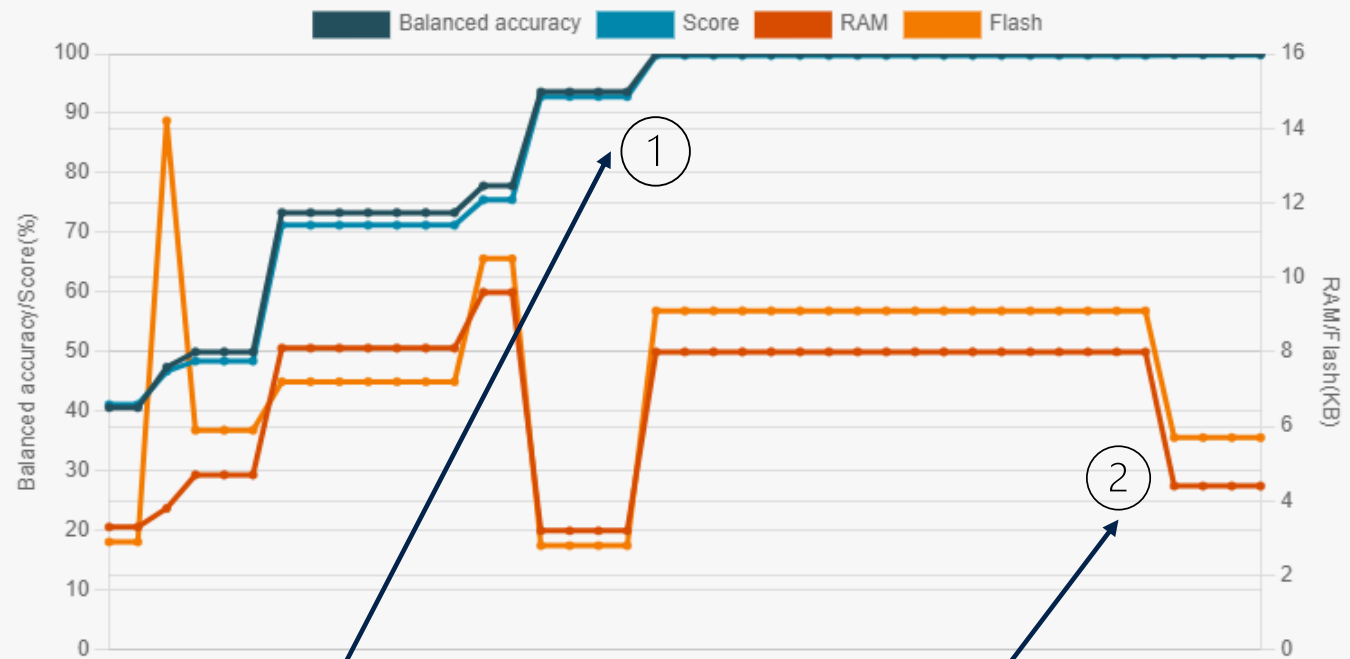
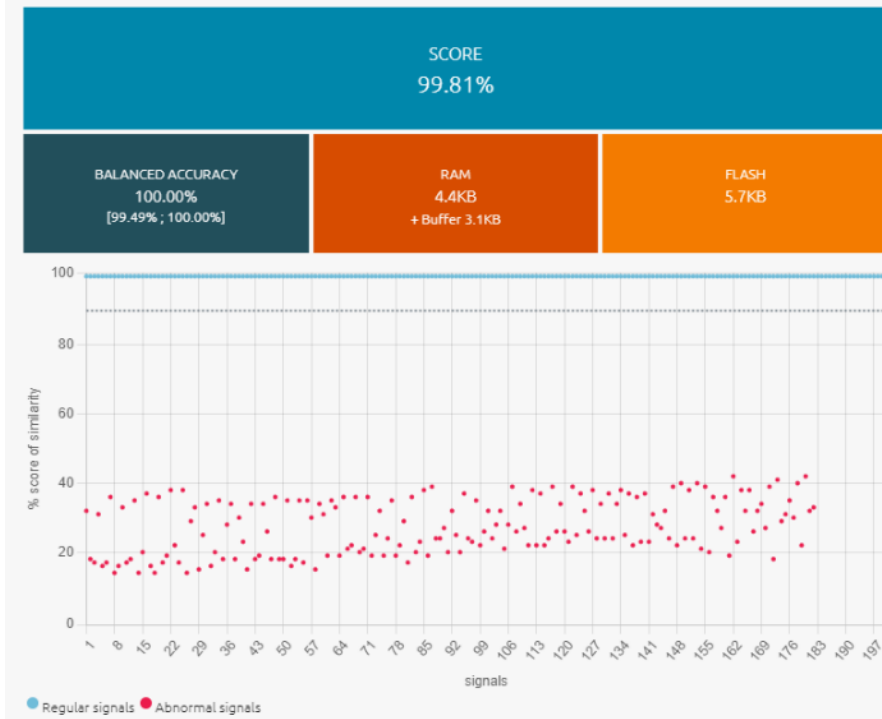
Below the File panel, there is an "Add file(s)" section with a dashed box and the text "Drop files or click to import".

1	2	3	4	5	6	7	8
-453	-1523	15386	517	-154	16106	1068	729
659	95	16368	1006	388	16779	313	-164
722	1012	15909	552	1007	17633	-461	-510
272	269	17697	-851	-911	17448	-836	-1172
45	-463	17629	-727	-1399	16834	-349	-1244

1.234	1.234	1.234
1.234	1.234	1.234
1.234	1.234	1.234
1.234	1.234	1.234
1.234	1.234	1.234
1.234	1.234	1.234

Get the best ML algorithm for your application

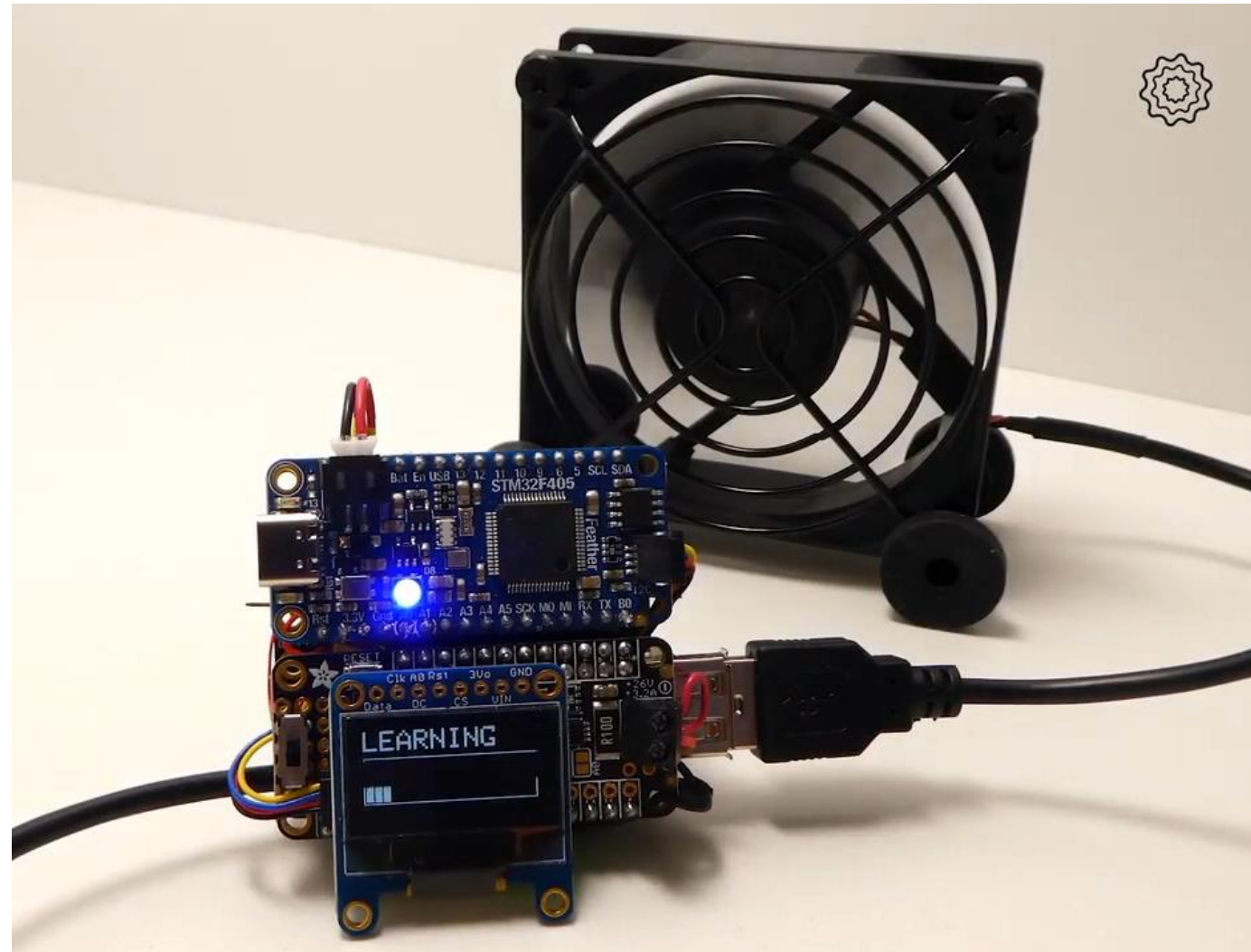
Internal benchmarking tool picks the best algorithm for your data



NanoEdge AI Studio improves the performance of the model..

.. And then optimizes it to reduce footprint and latency

On-device learning to learn on the edge and adapt to real environment



It's time to convert your data into valuable information!



Get inspired with our large library of case-studies



Download your free copy of NanoEdge AI Studio



<https://stm32ai.st.com>

Our technology starts with You



Find out more at www.st.com

© STMicroelectronics - All rights reserved.

ST logo is a trademark or a registered trademark of STMicroelectronics International NV or its affiliates in the EU and/or other countries.

For additional information about ST trademarks, please refer to www.st.com/trademarks.

All other product or service names are the property of their respective owners.



life.augmented