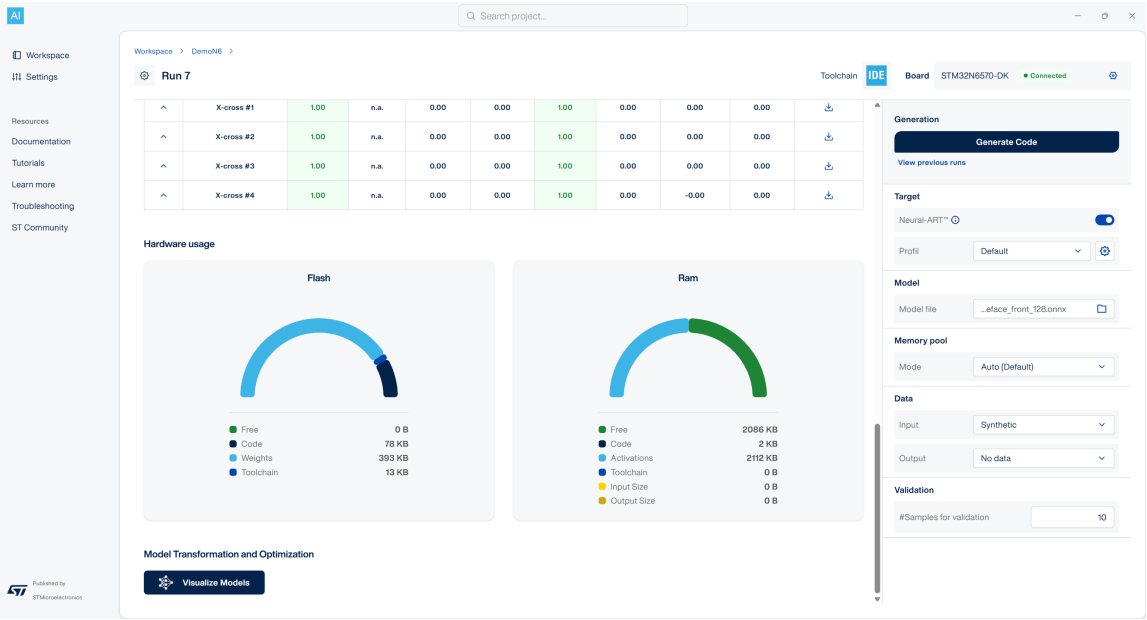
STM32Cube AI Studio: Compilation and optimization desktop tool for Edge AI on STM32 devices



DT74973V1

Product status
STEDGEAI-CUBEAI

ST Edge AI Core

STM32 Cube AI Studio

Features

- Generation of an STM32-optimized library from pre-trained neural network (NN) and classical machine learning (ML) models
- Provides detailed information about artificial intelligence (AI) model RAM and flash memory sizes
- Validates optimized models against reference models on the host and on the target
- Benchmark model performances on STM32 local board
- Support for STMicroelectronics Neural-ART Accelerator neural processing unit (NPU) for *AI*/*ML* model acceleration in hardware
- Native support for various deep learning frameworks such as Keras, TensorFlow™ Lite, LiteRT, and support for all frameworks that can export to the ONNX standard format such as PyTorch™, MATLAB®, and more
- Support for various built-in scikit-learn models such as isolation forest, support vector machine (SVM), and K-means via ONNX
- Easy portability across different STM32 microcontroller series through STM32Cube ecosystem compatibility
- Free-of-charge, user-friendly license terms

Description

STM32Cube *AI* Studio (STEDGEAI-CUBEAI) is STMicroelectronics desktop tool designed to evaluate, optimize and compile neural network (NN) models for STM32 microcontrollers. It also manages compilation of *NN* models for Neural-ART Accelerator neural processing unit (NPU). It replaces the X-CUBE-AI in the ST *AI* product offering to cover new STM32 devices.

STM32Cube *AI* Studio is a free-of-charge GUI allowing an automatic conversion of pretrained artificial intelligence algorithms, including neural network (NN) and classical machine learning models (ML), into the equivalent optimized C code to be embedded in the application. The generated optimized library offers an easy-to-use and developer-friendly method to deploy *AI* on edge devices. When optimizing *NN* models for Neural-ART Accelerator *NPU* accelerator, the tool generates the microcode that maps *AI* operations on the *NPU* when possible and falls back on the *CPU* when not.

STM32Cube *AI* Studio uses the ST Edge *AI* Core technology, which is STMicroelectronics technology (CLI) to optimize *NN* models for any STMicroelectronics products with *AI* capabilities.

Note: Neural-ART Accelerator is ST proprietary AI accelerator, embedded in some products, like the STM32N6 series. Refer to the datasheet if Neural-ART Accelerator is embedded in the targeted product.

ST Edge AI Suite

All the tools and software packages contributing to the STM32N6 AI ecosystem are part of STMicroelectronics ST Edge AI Suite, which is an integrated collection of software tools designed to facilitate the development and deployment of embedded AI applications. This comprehensive suite supports both optimization and deployment of machine learning algorithms and neural network models, from data collection to the final deployment on hardware, streamlining the workflow for professionals across various disciplines.

The ST Edge AI Suite supports various STMicroelectronics products: STM32 microcontrollers and microprocessors, Neural-ART Accelerator, Stellar microcontrollers, and smart sensors.

The ST Edge AI Suite represents a strategic move to democratize edge AI technology, making it a pivotal resource for developers looking to harness the power of AI in embedded systems efficiently and effectively.

1 General information

This document applies to the STM32 Arm® Cortex®-based microcontrollers.

arm

Note: Arm and Cortex are registered trademarks of Arm Limited (or its subsidiaries or affiliates) in the US and/or elsewhere.

The Arm word and logo are trademarks of Arm Limited (or its subsidiaries) in the US and/or elsewhere. All rights reserved.

TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

PyTorch is a trademark of Facebook, Inc.

1.1 Ordering information

STEDGEAI-CUBEAI is available for free download from the *www.st.com* website.

1.2 What is STM32Cube?

STM32Cube is an STMicroelectronics original initiative to improve designer productivity significantly by reducing development effort, time, and cost. STM32Cube covers the whole STM32 portfolio.

STM32Cube includes:

- A set of user-friendly software development tools to cover project development from conception to realization, among which are:
  - STM32CubeMX, a graphical software configuration tool that allows the automatic generation of C initialization code using graphical wizards
  - STM32CubeIDE, an Eclipse®-based IDE, providing code edition, compilation, programming, and debugging capabilities
  - STM32CubeCLT, an all-in-one command-line development toolset with code compilation, board programming, and debug features
  - STM32CubeIDE for Visual Studio Code (STM32VSCode), a complete IDE based on VS Code® platform
  - STM32CubeProgrammer (STM32CubeProg), a programming tool available in graphical and command-line versions
  - STM32CubeMonitor (STM32CubeMonitor, STM32CubeMonPwr, STM32CubeMonRF, STM32CubeMonUCPD), powerful monitoring tools to fine-tune the behavior and performance of STM32 applications in real time
  - STM32CubeWiSE (STM32CubeWiSEcg, STM32CubeWiSEre), graphical tools designed to evaluate and test the capabilities of sub-GHz radios and protocols
- STM32Cube MCU and MPU Packages, comprehensive embedded-software platforms specific to each microcontroller and microprocessor series (such as STM32CubeWBA for the STM32WBA series), which include:
  - STM32Cube hardware abstraction layer (HAL), ensuring maximized portability across the STM32 portfolio
  - STM32Cube low-layer APIs, ensuring the best performance and footprints with a high degree of user control over hardware
  - A consistent set of middleware components such as ThreadX, FileX, LevelX, NetX Duo, FreeRTOS™, USBX, touch library, mbed-crypto, TFM, MCUboot, OpenBL, and STM32_WPAN (including Bluetooth® LE profiles and services, Mesh, Zigbee®, OpenThread, Matter, and 802.15.4 MAC layer)
  - All embedded software utilities with full sets of peripheral and applicative examples
- STM32Cube Expansion Packages, which contain embedded software components that complement the functionalities of the STM32Cube MCU and MPU Packages with:
  - Middleware extensions and applicative layers
  - Examples running on some specific STMicroelectronics development boards

2 Key benefits

- **Fast, automated edge *AI* deployment**
  Converts pre-trained neural network (NN) and classical machine learning (ML) models into optimized C libraries for STM32, dramatically reducing manual porting and integration efforts.
- **Maximized performance and efficiency on STM32**
  Provides detailed RAM/flash usage and optimizes/schedules operations on the Neural-ART Accelerator *NPU* (with *CPU* fallback).
- **Seamless integration in the STM32 ecosystem**
  Free desktop tool for Windows® and Linux®, aligned with the STM32Cube ecosystem. This enables easy portability across STM32 series.

3 License

STEDGEAI-CUBEAI is delivered under the SLA0048 software license agreement and its Additional License Terms.

Revision history

Table 1. **Document revision history**

Date	Version	Changes
16-Feb-2026	1	Initial release.

Glossary

AI Artificial intelligence

CPU Central processing unit

ML Machine learning

NN Neural network

NPU Neural processing unit