Embedded AI and TinyML: running models on microcontrollers

26 October 2026 · 5 min read · blogs.iictp.org/tinyml-embedded-ai

TinyML runs machine-learning models on microcontrollers with very little memory and power. It suits simple, always-on tasks such as detecting a keyword, a vibration pattern or an anomaly in a sensor stream.

TL;DR The short version

- TinyML runs machine-learning models on microcontrollers with very little memory and power.
- It suits simple always-on tasks such as keyword, vibration or anomaly detection, responding instantly and running long on a battery.
- Fit the model with a small architecture for one narrow task and 8-bit quantization, and accept limited capability.

Read online, listen to it, or change the text size: https://blogs.iictp.org/tinyml-embedded-ai

Why do it

- Responds immediately, without a network.
- Uses very little power, so it can run on a battery for a long time.
- Raw data, such as audio, never leaves the device.

Making a model fit

- Choose a small architecture for one narrow task.
- Quantize the model to 8-bit integers.
- Convert it with a runtime built for microcontrollers, such as TensorFlow Lite for Microcontrollers.
- Measure memory, speed and energy on the real chip.

Limits

Models are small and capabilities limited. Updating them in the field is harder than updating a cloud service, so plan for it. Test on the real sensor data and in the real environment, since a model can behave differently away from the lab.

Related credential: IoT & Embedded Systems Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. TensorFlow Lite for Microcontrollers. tensorflow.org/lite/microcontrollers.

Published by IICTP, an independent technical credential body, on 26 October 2026. The current version of this post is at https://blogs.iictp.org/tinyml-embedded-ai. More guides: blogs.iictp.org.