# 302. BRINGING ADVANCED AI TO MICROCONTROLLERS

*Shved I.A.*
*Belarusian State University of Informatics and Radioelectronics*
*Minsk, Republic of Belarus*

*Kaspiarovich N.G. – Senior Lecturer*

The deployment of advanced Artificial Intelligence on resource-constrained microcontrollers - a field known as TinyML - is evolutionizing edge computing. This paper explores the critical need to move computations from the cloud to ultra-low-power devices to enable real-time, privacy-preserving, and energy-efficient applications.

Nowadays IoT is expanding and developing rapidly. We already have our everyday tasks automated or at least facilitated by the help of smart devices. A lot of data is generated by different IoT devices: from sensors and wearables to industrial monitors. Traditionally this data is streamed to the cloud, where powerful servers run AI models to extract insights. While effective, this cloud-centric model suffers from inherent drawbacks: high latency for time-sensitive applications, significant bandwidth costs, and potential privacy risks associated with transmitting personal or sensitive data. To evade these limitations we are shifting towards edge AI, where intelligence is brought directly on the devices that collect the data.

At the forefront of this shift is TinyML, one of the fastest-growing areas of Deep Learning nowadays. TinyML is a field dedicated to deploying AI models on embedded devices with extremely low power consumption, often in the milliwatt range or lower. By integrating embedded systems and AI, TinyML facilitates local data processing on edge devices, reducing dependency on cloud computing while enhancing data privacy, decreasing latency, and minimizing power consumption [1]. It should be taken into account that according to ABI Research forecast [2], by 2030 about 2.5 billion devices will enter the market using TinyML technologies.

However, there are some challenges that this area is facing. They are the compute-memory gap and the power and energy constraints. To solve them a suite of optimization techniques has been developed.

Quantization. The division of a quantity into a discrete number of small parts is often assumed to be integral multiples of a common quantity. In general, it is the mapping of a continuous quantity into a discrete one [3].

Pruning. Pruning is a technique inspired by neurobiology that removes redundant or less important connections (weights) from a neural network. After training, weights with magnitudes close to zero contribute little to the final output and can be removed. This creates a sparse model, reducing the number of parameters to store and the number of operations to compute.

Knowledge Distillation. Knowledge distillation is a model compression technique that trains a smaller, simpler model to mimic the behavior of a larger, more complex model.

The optimization techniques described above are operationalized by a rapidly growing ecosystem of software frameworks and hardware innovations.

From the software side we already have many frameworks that allow you to train your own lightweight model that is ready to be deployed, for example, TensorFlow Lite for Microcontrollers. It is an open-source deep-learning framework that has a set of tools providing on-device machine learning. There are already various datasets and pretrained models, which simplify the process.

Hardware Advancements: MCUs with AI Accelerators. The most recent and impactful advancement is the integration of specialized hardware for neural network acceleration directly into MCUs. These are often referred to as microNPUs (Neural Processing Units). They are optimized to perform the massive parallel computations of matrix multiplications and convolutions with exceptional energy efficiency, often improving performance by orders of magnitude compared to running the same operations on the CPU alone.

The research shows that bringing AI to the microcontrollers is not only feasible but is a rapidly maturing field. Looking forward, the field is poised for continued, rapid evolution. Future investigation will likely focus on enabling more sophisticated on-device learning (lifelong or continuous learning) to allow devices to adapt to new data without cloud connectivity. Ultimately, the trajectory points towards a world with trillions of intelligent, sensing, and acting devices, creating an ambient intelligence that will fundamentally reshape our interaction with the physical world.

***References:***

*1. Sanchez-Iborra, R. TinyML-enabled frugal smart objects: challenges and opportunities / R. Sanchez-Iborra, A. F. Skarmeta // IEEE Circuits and Systems Magazine. – 2020. – Vol. 20, № 3. – P. 4–18. – URL: https://ieeexplore.ieee.org/document/9166461 (date of access: 03.03.2026).*

*2. Global Shipments of TinyML Devices to Reach 2.5 Billion by 2030: [website]. – USA, 2026. – URL: https://www.abiresearch.com/press/global-shipments-tinyml-devices-reach-25-billion-2030 (date of access: 03.03.2026).*

*3. Gray, R. M. Quantization / R. M. Gray, D. L. Neuhoff // IEEE Transactions on Information Theory. – 1998. – Vol. 44, № 6. – P. 2325–2383. – URL: https://www.math.ucdavis.edu/~saito/data/quantization/44it06-gray.pdf (date of access: 03.03.2026).*