Elevating AI on the Edge: A Demonstration of MIMaaS (Machine Intelligence with Microcontroller-as-a-Service)

Sebastian Zimmermann, René Groh, Andreas M. Kist

Department Artificial Intelligence in Biomedical Engineering (AIBE)
Friedrich-Alexander-Universität Erlangen-Nürnberg
Nürnberger Straße 74, 91052 Erlangen, Germany

Abstract

Deploying AI on microcontrollers (MCUs) is challenging. We introduce *MIMaaS*, a Microcontroller-as-a-Service platform that enables users to upload a model, select a target device, and receive a detailed performance report remotely. A key innovation is our measurement of real-world power consumption, alongside latency and memory usage, directly from the physical hardware. MIMaaS empowers researchers and developers to easily create and validate hardware-aware AI models without needing physical hardware access.

Introduction

The field of Edge Artificial Intelligence (Edge AI) is rapidly expanding, bringing sophisticated data processing directly to billions of resource-constrained edge devices (Singh and Gill 2023; Abadade et al. 2023; Yao et al. 2023). This trend, however, often clashes with the high demands of hardware deployment and testing. Validating AI models across the wide landscape of available MCUs presents financial and logistical hurdles. Procuring a diverse collection of hardware is prohibitively expensive for many developers and researchers. Furthermore, the manual setup and configuration of physical testbeds are both complex and time-consuming. This accessibility barrier stifles innovation, making it difficult to benchmark performance on various platforms without a substantial upfront investment.

To bridge this gap, we present *MIMaaS*, a novel Microcontroller-as-a-Service platform. Our service empowers users to thoroughly evaluate their neural network models on a desired hardware platform via REST-API or a website. Models will first be checked for general compatibility with the desired platform. Given compatibility, the model is deployed on the hardware and users will receive a detailed performance analysis, including inference time, memory usage (RAM and ROM), and a fine-grained power consumption profile, which is our key innovation. This massively lowers the threshold for creating and validating hardware-aware AI models without requiring physical hardware access.

System Architecture

The system architecture is composed of three core components: a flexible model submission interface, a scalable processing backend, and a comprehensive performance/error analysis report (see Figure 1).

Submission Interface: The system offers two model submission interfaces to accommodate different user workflows. A web-based graphical user interface allows for the submission of individual models via drag-and-drop. For automated processes, a Python REST API is provided to facilitate programmatic access, enabling seamless integration into larger workflows such as Neural Architecture Search and other AutoML pipelines.

Server-Sided Processing and Profiling: The backend consists of a multi-core workstation which directs incoming jobs to a cluster of physical MCUs. To ensure high throughput and scalability, the system parallelizes the compilation and deployment, allowing multiple models and targets to be evaluated concurrently. Each device is connected to a high-fidelity power monitor (Power Profiler Kit 2 by Nordic Semiconductor). In our typical current range of 500 μA - 5 mA the PPK2 measures the power consumption with an accuracy of $\pm$10% and offset of $\pm$2% (Nordic 2025b) at a sampling rate of 100 kHz (Nordic 2025a). Unfortunately, Nordic does not provide exact information about the utilized current sensor. From the raw data, we extract key metrics, including precise inference time and the total energy consumed per inference cycle. (See Figure 2)

Analysis and Results: Upon completion of the profiling job, the user receives a comprehensive performance report. The key results include on-device metrics such as memory footprint (ROM and RAM), precise inference latency, and total energy consumption. Additionally, the report contains the fine-grained power consumption profile, offering detailed insight into the model's behavior on the target hardware.

System Costs: The overall system costs primarily depend on the number of measurement nodes and the workstation configuration. Each node contains of an MCU and a dedicated PPK2, which results in approximately €100 per node. The multi-core workstation used for compilation and job scheduling costs approximately €900.

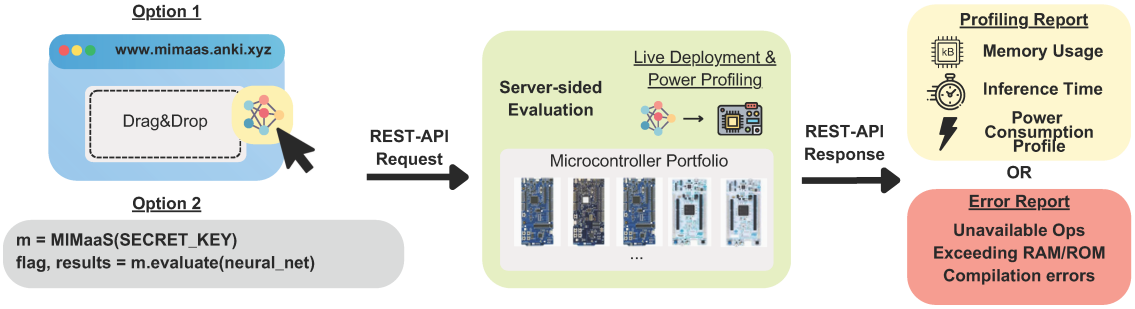

Figure 1: Overview of the MIMaaS system pipeline.

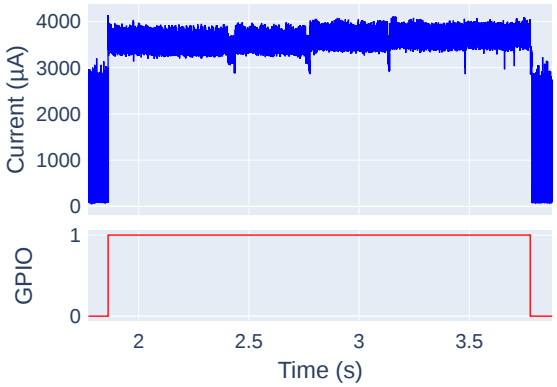

Figure 2: Example current draw measured during model inference.

Brand	Board
Nordic Semiconductor	nRF52840-DK
	nRF5340-DK
	nRF52833-DK
STMicroelectronics	STM32F4 Discovery (WIP)
	STM32L4 Nucleo (WIP)
Espressif Systems	ESP32-DevKitC (WIP)

Table 1: Portfolio of supported MCUs or work in progress (WIP) (As of November 2025). New MCUs are constantly added to the MIMaaS hardware library.

The Demonstration Experience

The demonstration guides a user through a streamlined, four-step workflow from model selection to performance analysis.

1. Model Selection: Participants can either train a custom neural network or select one from a curated set of pre-trained models, all using a provided Jupyter Notebook on Google Colab.

2. Target Configuration: Next, users choose a target MCU from a list of available hardware platforms (see Table 1) and specify key deployment parameters, such as quantization.

3. One-Click Deployment: With a single click, the configured job is submitted to a server queue. It is then automatically checked for compatibility, then compiled, deployed, and profiled on the physical hardware target.

4. Performance Analysis: Upon completion, the system returns a detailed report. A successful run provides the metrics as stated in the previous section. If an error occurs, the system provides a diagnostic message and guidance for resolving the issue, including constraints in neural network architecture or size limitations.

Novelty and Innovation

Several cloud-based platforms, such as Edge Impulse (Hymel et al. 2022), allow developers to deploy and evaluate AI models on MCUs. However, these systems do not perform direct measurements of the power consumption of the device under test (DUT). Instead, they provide estimated values whose derivation is not publicly documented. In contrast, the proposed *MIMaaS* platform introduces a fully open-source deployment and profiling pipeline that seamlessly integrates into the development workflow. It provides detailed feedback on hardware-specific model characteristics (RAM/ROM usage and inference time) and performs real-time measurement of the DUT's power consumption during inference — a feature particularly valuable for developers of energy-constrained devices such as wearables. To ensure transparency and enable community-driven extensions, we plan to release MIMaaS under a permissive open-source license (CC BY-NC-SA). For convenience, we also provide a hosted for-profit version of the system.

Conclusion and Future Work

We presented MIMaaS, an open-source platform that simplifies profiling AI models on physical MCUs in the cloud. It delivers key metrics including latency, memory footprint, and, most notably, real-world power consumption, allowing for rapid, hardware-aware development. Future work will focus on expanding our hardware library beyond general MCUs to specific NPUs (Chen et al. 2020), FPGAs (Kalapothas, Flamis, and Kitsos 2022), and AI frameworks, such as PyTorch, and deepening integration with AutoML pipelines (Groh and Kist 2023; Liberis, Dudziak, and Lane 2021). A further increased measurement accuracy could be achieved with custom-designed measurement devices (Aragon Jurado et al. 2025).

References

Abadade, Y.; Temouden, A.; Bamoumen, H.; Benamar, N.; Chtouki, Y.; and Hafid, A. S. 2023. A comprehensive survey on tinyml. *IEEE Access*, 11: 96892–96922.

Aragon Jurado, J.; de la Torre Macías, J.; Ruiz, P.; and Dorronsoro, B. 2025. Automatic software tailoring for Green Internet of Things. *Internet of Things*, 30: 101521.

Chen, Y.; Xie, Y.; Song, L.; Chen, F.; and Tang, T. 2020. A survey of accelerator architectures for deep neural networks. *Engineering*, 6(3): 264–274.

Groh, R.; and Kist, A. M. 2023. End-to-end evolutionary neural architecture search for microcontroller units. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, 1–7. IEEE.

Hymel, S.; Banbury, C.; Situnayake, D.; Elium, A.; Ward, C.; Kelcey, M.; Baaijens, M.; Majchrzycki, M.; Plunkett, J.; Tischler, D.; et al. 2022. Edge impulse: An mlops platform for tiny machine learning. *arXiv preprint arXiv:2212.03332*.

Kalapothas, S.; Flamis, G.; and Kitsos, P. 2022. Efficient edge-AI application deployment for FPGAs. *Information*, 13(6): 279.

Liberis, E.; Dudziak, Ł.; and Lane, N. D. 2021. μnas: Constrained neural architecture search for microcontrollers. In *Proceedings of the 1st Workshop on Machine Learning and Systems*, 70–79.

Nordic. 2025a. Power Profiler Kit II Brief Product Information. https://nsscprodmedia.blob.core.windows.net/prod/software-and-other-downloads/product-briefs/power-profiler-kit-ii-pbv10.pdf. Accessed: November 12, 2025.

Nordic. 2025b. Power Profiler Kit II Documentation. https://docs.nordicsemi.com/bundle/ug_ppk2/page/UG/ppk/ppk_measure_resolution.html. Accessed: November 12, 2025.

Singh, R.; and Gill, S. S. 2023. Edge AI: a survey. *Internet of Things and Cyber-Physical Systems*, 3: 71–92.

Yao, J.; Zhang, S.; Yao, Y.; Wang, F.; Ma, J.; Zhang, J.; Chu, Y.; Ji, L.; Jia, K.; Shen, T.; Wu, A.; Zhang, F.; Tan, Z.; Kuang, K.; Wu, C.; Wu, F.; Zhou, J.; and Yang, H. 2023. Edge-Cloud Polarization and Collaboration: A Comprehensive Survey for AI. *IEEE Transactions on Knowledge and Data Engineering*, 35(7): 6866–6886.