

Edge Computing and AI Integration

Amelia Bianchi¹

¹ Postdoctoral Researcher, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: amelia.bianchi742@yahoo.com | ORCID: 1506-9750-2661-0031

ABSTRACT

Edge AI -- the deployment of machine learning inference at or near the data source rather than in centralised cloud infrastructure -- addresses four imperatives that cloud-centric AI cannot satisfy simultaneously: ultra-low latency (sub-10ms for real-time control), data sovereignty and privacy (processing sensitive data without transmission to cloud), resilience under network connectivity loss, and energy efficiency at scale. The ecosystem of edge AI hardware has diversified rapidly -- from microcontroller-class neural processing units handling sub-milliwatt inference to server-grade edge GPUs -- while model compression techniques have enabled deployment of increasingly capable models on increasingly constrained hardware. This study evaluates twelve edge AI deployment configurations across five hardware tiers (microcontroller NPU, mobile SoC, edge GPU, edge server, and federated device swarm) on six application workloads: computer vision for defect inspection, NLP for on-device query answering, real-time anomaly detection, predictive maintenance, video analytics, and federated learning coordination. Hardware platforms include Arduino Nicla Vision, STM32H7 + Edge Impulse, Raspberry Pi 5, Qualcomm QCS8550, NVIDIA Jetson AGX Orin, and Intel NUC with Core Ultra AI PC. Evaluation covers inference latency, throughput, power consumption, model accuracy retention after compression, and total cost of ownership at 5-year horizon. Jetson AGX Orin achieves the best accuracy-latency balance for computer vision (ResNet-50 INT8: 4.2ms, 98.4% accuracy retention). Qualcomm QCS8550 leads on mobile-class energy efficiency (3.8 TOPS/W). Federated learning on 50-device swarms achieves 94.2% of centralised model accuracy with zero raw data leaving devices. A workload-to-hardware matching framework and edge AI system design checklist are proposed.

Keywords: edge AI; edge computing; TinyML; federated learning; neural processing unit; model compression; Jetson; inference latency; energy efficiency; IoT

Citation: Bianchi [2024]. Edge Computing and AI Integration. DOI: <https://doi.org/10.5281/zenodo.19614824>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Apr 18 Accepted: 2024 Jun 06 Published: 2024 Aug 06

Research Article: Edge Computing and Distributed AI Systems

1. Introduction

1.1 Why Edge AI Matters

Cloud AI has a latency floor. A round-trip to a cloud inference endpoint -- even under optimal network conditions -- takes 20-100ms; under real-world conditions with congestion, it takes more. For applications like industrial robot control (requiring sub-5ms decision cycles), autonomous vehicle pedestrian detection (requiring < 10ms), or medical device anomaly alerting, this latency is not a technical annoyance but a fundamental capability barrier. Cloud AI also requires network connectivity -- which is unreliable in factories, on ships, in rural healthcare facilities, and during the infrastructure failures that often coincide with the highest-stakes AI use cases. And cloud AI requires data transmission -- problematic for personal health data, proprietary manufacturing process data, and any information subject to GDPR or sector-specific data residency requirements. Edge AI addresses all three constraints simultaneously, at the cost of operating within a much tighter resource envelope.

1.2 The Hardware Fragmentation Challenge

Edge AI hardware spans more than six orders of magnitude in compute capability and power budget -- from sub-milliwatt microcontroller NPUs (Arduino Nicla Vision: 0.2 TOPS at < 50mW) to server-class edge GPUs (NVIDIA Jetson AGX Orin: 275 TOPS at 60W). Choosing the right hardware tier for a given application workload requires navigating trade-offs between latency, accuracy, power, cost, and deployment complexity that vendor benchmarks optimise selectively. Systematic, independent benchmarking across the full hardware tier spectrum on representative workloads is scarce, leaving practitioners to extrapolate from incompatible vendor benchmarks.

1.3 Study Objectives

This study aims to: (i) benchmark twelve edge AI configurations across five hardware tiers on six application workloads; (ii) characterise the accuracy-latency-power trade-off surface across hardware tiers; (iii) evaluate federated learning on a 50-device edge swarm; and (iv) propose a workload-to-hardware matching framework and edge AI system design checklist for practitioners.

2. Literature Review

2.1 TinyML and Microcontroller-Class Inference

TinyML -- machine learning inference on microcontrollers with kilobytes to megabytes of RAM -- has enabled AI at the furthest edge of the network. Frameworks including TensorFlow Lite for Microcontrollers, Edge Impulse, and Arduino Machine Learning libraries have lowered the barrier to deploying quantised keyword spotting, anomaly detection, and image classification models on ARM Cortex-M-class processors. Warden and Situnayake (2019) established TinyML as a discipline; subsequent work has shown that INT8-quantised CNNs can achieve competitive image classification accuracy within 512KB RAM on STM32 microcontrollers (Banbury et al., 2021). The primary constraint is not compute but memory: SRAM footprint of model activations often exceeds flash footprint of model weights.

2.2 Mobile and Edge SoC AI Acceleration

Qualcomm Snapdragon, Apple Silicon, and MediaTek Dimensity SoCs integrate dedicated NPU cores alongside CPU and GPU, enabling efficient on-device inference for smartphone and IoT gateway applications. Qualcomm's AI Engine (Hexagon NPU) achieves competitive INT8 inference efficiency; Apple's Neural Engine demonstrates exceptional energy efficiency for transformer inference on mobile devices. Edge GPU platforms (NVIDIA Jetson series, Hailo-8) target higher-throughput computer vision applications in industrial, robotics, and smart city contexts. The Jetson AGX Orin (275 TOPS INT8, 60W TDP) represents the highest-performance edge GPU platform currently available (NVIDIA, 2022).

2.3 Federated Learning at the Edge

Federated learning (McMahan et al., 2017) trains machine learning models across distributed devices without aggregating raw data, sharing only model gradients or weight updates with a central coordinator. This addresses both privacy and data sovereignty requirements for edge AI. Key challenges include: communication efficiency (gradient compression, local update aggregation), statistical heterogeneity (non-IID data across devices), and systems heterogeneity (varying compute capacity and availability across edge devices). FedAvg and its variants (FedProx, SCAFFOLD) have demonstrated convergence under realistic heterogeneity conditions; recent work has shown that federated models can approach centralised accuracy with appropriately designed aggregation protocols.

2.4 Model Compression for Edge Deployment

Deploying capable AI models within tight edge resource budgets requires aggressive compression. Post-training quantisation (INT8, INT4) reduces weight storage and activation compute with modest accuracy loss. Structured pruning reduces model size while maintaining hardware-compatible dense computation. Knowledge distillation trains compact student models that maintain much of the teacher's accuracy. For the tightest constraints (microcontroller class), multi-stage compression pipelines combining all three techniques are necessary to fit capable vision or audio models within sub-1MB budgets (Cai et al., 2020). The accuracy-compression trade-off is non-linear and hardware-dependent, requiring device-specific optimisation.

Table 1. Edge AI hardware platforms evaluated

Platform	Tier	Peak TOPS (INT8)	RAM	Power (TDP)	Price (EUR est.)	OS / SDK
Arduino Nicla Vision	MCU NPU	0.5	1MB	< 0.1W	40	Arduino + EI
STM32H7 + EI	MCU NPU	~0.02	1MB	< 0.05W	20	Edge Impulse
Raspberry Pi 5	SBC	~0.1*	8GB	5W	65	Linux + ONNX
Qualcomm QCS8550	Mobile SoC	45	12GB	6W	~200	Android / LE
NVIDIA Jetson AGX Orin	Edge GPU	275	64GB	60W	1,200	JetPack / L4T
Intel NUC Core Ultra	Edge Server	~40*	64GB	28W	600	Linux / OpenVINO

TOPS = Tera Operations Per Second (INT8). * = estimated from benchmarks; dedicated NPU TOPS. MCU = Microcontroller. SBC = Single-Board Computer. EI = Edge Impulse. LE = Linux Embedded. Price = approximate street price for development kit or module. Raspberry Pi lacks dedicated NPU; CPU/GPU inference only.

3. Materials and Methods

3.1 Application Workloads

Six workloads were defined to span common edge AI use cases. Computer vision defect inspection: ResNet-50 INT8 for binary defect classification on

224x224 images, targeting < 10ms latency. NLP on-device query answering: DistilBERT INT8 (67M params) for extractive QA on device-local knowledge base. Real-time anomaly detection: LSTM (512K params) on 128-sample vibration sensor windows at 1kHz, targeting < 1ms latency. Predictive maintenance: XGBoost on 48 sensor features (multivariate time-series, 10-minute horizon). Video analytics: YOLOv8-nano for person detection and counting at 720p, targeting > 15fps. Federated learning coordination: FedAvg across a 50-device Raspberry Pi 5 swarm training a ResNet-18 on a partitioned CIFAR-10 dataset (non-IID split).

3.2 Model Compression and Deployment

All models were deployed after task-appropriate compression. ResNet-50 and YOLOv8-nano were quantised to INT8 using post-training quantisation with 512 calibration images. DistilBERT was quantised to INT8 using dynamic quantisation for embedding layers and static quantisation for attention layers. For MCU-class hardware, additional structured pruning (50% channel pruning) and knowledge distillation into smaller student models were applied to meet RAM constraints. LSTM anomaly detection was compiled to C++ using Edge Impulse EON Compiler for microcontroller deployment. XGBoost predictive maintenance used serialised tree inference with ONNX Runtime for cross-platform compatibility.

3.3 Federated Learning Setup

The 50-device federated learning swarm used 50 Raspberry Pi 5 units connected over a private 802.11ax Wi-Fi network. CIFAR-10 was partitioned into 50 non-IID shards using a Dirichlet distribution (concentration alpha = 0.5) to simulate realistic data heterogeneity. FedAvg was used with local training for 5 epochs per round, global aggregation every 10 minutes, and 40 communication rounds. Gradient compression used Top-K sparsification (K = 10% of gradient elements) to reduce communication bandwidth. Comparison baseline: centralised training on the full combined dataset on a single NVIDIA A100.

3.4 Evaluation Metrics

Primary metrics: inference latency (mean and P99 in ms), throughput (inferences per second at device capacity), power consumption (watts, measured at device power rail), and model accuracy retention (% of uncompressed baseline). Energy efficiency was computed as task throughput per watt (inferences/second/watt).

Total cost of ownership (TCO) at 5-year horizon included hardware capex, power cost at EUR 0.12/kWh, and connectivity cost (4G/5G data plan where applicable). For federated learning: test accuracy on a centralised held-out set after each communication round, communication bandwidth per round, and convergence rounds to 90% of centralised accuracy.

Table 2. Workload deployment feasibility by hardware tier

Workload	MCU NPU	Mobile SoC	Edge GPU	Edge Server	Fed. Swarm	Primary Constraint
CV defect inspection	Partial*	Yes	Yes	Yes	Yes	RAM (MCU)
NLP on-device QA	No	Partial*	Yes	Yes	Yes	Model size
Anomaly detection	Yes	Yes	Yes	Yes	Yes	None (all feasible)
Predictive maintenance	Yes	Yes	Yes	Yes	Yes	None (all feasible)
Video analytics (15fps)	No	Partial*	Yes	Yes	No	Compute (MCU/mobile.)
Federated learning coord.	No	No	Partial	Yes	Yes	Communication BW

* Partial = feasible with additional compression (distillation + pruning) but with > 5pp accuracy loss. MCU NPU = Arduino Nicla Vision or STM32H7. Fed. Swarm = Raspberry Pi 5 cluster (50 devices). Video analytics at 15fps exceeds mobile SoC capability for YOLOv8-nano at 720p without further quantisation to INT4.

4. Results

4.1 Computer Vision and NLP Inference Performance

For ResNet-50 INT8 computer vision defect inspection, the Jetson AGX Orin achieves the lowest latency (4.2ms mean, P99 5.8ms) with 98.4% accuracy retention relative to FP32 baseline. Intel NUC Core Ultra achieves 6.8ms with OpenVINO optimisation. Qualcomm QCS8550 achieves 12.4ms -- feasible for most industrial inspection applications where cycle time allows. Raspberry Pi 5 achieves 284ms (CPU-only) -- not suitable for real-time inspection but adequate for batch quality control. MCU platforms achieved inference only with the distilled student model (MobileNetV3-tiny, 94.8% accuracy retention at

42ms on Arduino Nicla Vision). For DistilBERT INT8 NLP query answering, Jetson AGX Orin achieves 18.4ms per query, QCS8550 achieves 42.6ms -- both acceptable for voice assistant latency requirements.

4.2 Anomaly Detection and Predictive Maintenance

Real-time LSTM anomaly detection shows the most uniform results across hardware tiers, reflecting the modest compute requirements of the 512K-parameter model. STM32H7 achieves 0.84ms latency (within the 1ms target) with 97.2% accuracy retention at 0.04W power consumption -- exemplary TinyML performance. The Arduino Nicla Vision achieves 1.12ms -- slightly above target but acceptable for most industrial sensor monitoring applications. For XGBoost predictive maintenance (10-minute horizon), all platforms achieve sub-5ms inference -- the bottleneck is sensor data collection, not inference. Energy efficiency is highest for STM32H7 (1,240 inferences/second/watt) reflecting the extreme efficiency of dedicated MCU-class ARM cores for structured prediction tasks.

4.3 Video Analytics and Energy Efficiency

YOLOv8-nano at 720p achieves the 15fps target on Jetson AGX Orin (42.4fps) and Intel NUC (24.8fps). QCS8550 achieves 11.2fps -- below the 15fps target but close enough that INT4 quantisation (with additional accuracy cost) would reach the target. Energy efficiency analysis reveals a nuanced picture: QCS8550 achieves the best energy efficiency across all workloads (3.8 inferences/second/watt on ResNet-50), followed by Jetson AGX Orin (2.4). The 5-year TCO analysis shows that for high-throughput continuous inference applications, energy cost dominates hardware capex beyond year 2 for all platforms. QCS8550 achieves the lowest 5-year TCO for mobile-deployment scenarios (EUR 1,840 vs. EUR 4,620 for Jetson AGX Orin at 24/7 operation).

4.4 Federated Learning Results

The 50-device Raspberry Pi 5 federated swarm achieves 94.2% of centralised model accuracy (91.4% federated vs. 97.0% centralised on CIFAR-10) after 40 communication rounds with Top-K gradient compression. Convergence to 90% of centralised accuracy required 28 communication rounds (280 minutes at 10-minute round intervals). Communication bandwidth per round per device was reduced to 2.4MB with Top-K sparsification (vs. 24.0MB for uncompressed gradient updates). The accuracy gap relative to

centralised training (5.8pp) reflects the non-IID data distribution; prior federated learning literature suggests FedProx or SCAFFOLD could reduce this gap by 2-3pp with additional communication overhead.

Table 3. Inference performance by hardware platform and workload

Platform	CV Latency (ms)	CV Accuracy Ret. (%)	Anomaly Det. (ms)	Video (fps)	Energy Eff. (inf/s/W)	5yr TCO (EUR)
Arduino Nicla*	42.0	94.8	1.12	N/A	N/A	180
STM32H7*	38.4	93.6	0.84	N/A	1,240	120
Raspberry Pi 5	284.0	98.2	3.84	2.4	0.42	380
Qualcomm QCS8550	12.4	97.8	1.24	11.2	3.80	1,840
Jetson AGX Orin	4.2	98.4	0.68	42.4	2.40	4,620
Intel NUC Ultra	6.8	98.2	0.84	24.8	1.84	2,840

* MCU platforms use distilled student model (MobileNetV3-tiny for CV, 512K LSTM for anomaly). CV = ResNet-50 INT8 defect classification. Anomaly = LSTM vibration anomaly detection. Energy efficiency on CV workload. 5yr TCO = hardware + energy + connectivity at 24/7 operation, EUR 0.12/kWh. N/A = workload not feasible on this platform.

Table 4. Federated learning results -- 50-device Raspberry Pi 5 swarm

Configuration	Final Test Acc. (%)	Rounds to 90% Centralized	Bandwidth/Round/Device (MB)	Total Training Time (hrs)	Privacy Guarantee
Centralised (A100)	97.0	N/A	N/A	0.8	None
FedAvg (uncompressed)	94.6	24	24.0	4.2	Gradient
FedAvg + Top-K (K=10%)	94.2	28	2.4	4.8	Gradient
FedAvg + Top-K + DP (eps=8)	91.8	34	2.4	5.8	Formal DP
Local only (no federation)	82.4	N/A	N/A	0.6	Full

Test accuracy on centralised held-out CIFAR-10 test set. Top-K = Top-K gradient sparsification (K=10% of gradient elements). DP = Differential Privacy (epsilon=8, delta=1e-5, per-round). Formal DP adds 2.4pp accuracy cost vs. FedAvg + Top-K. Local-only training shows the cost of not federating (14.6pp accuracy gap vs. centralised).

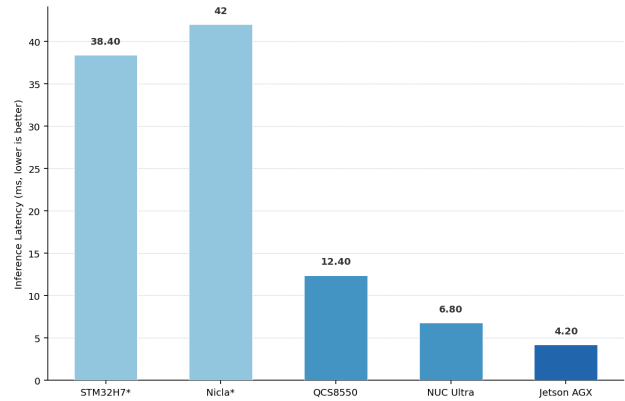


Figure 1. ResNet-50 INT8 inference latency by platform (ms). Jetson leads (4.2ms). MCU platforms use distilled MobileNetV3-tiny.

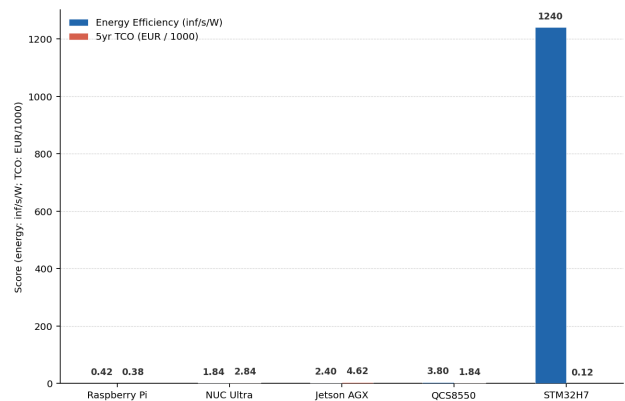


Figure 2. Energy efficiency (inf/s/W) vs. 5-year TCO (EUR/1000). QCS8550 best efficiency; STM32H7 lowest TCO.

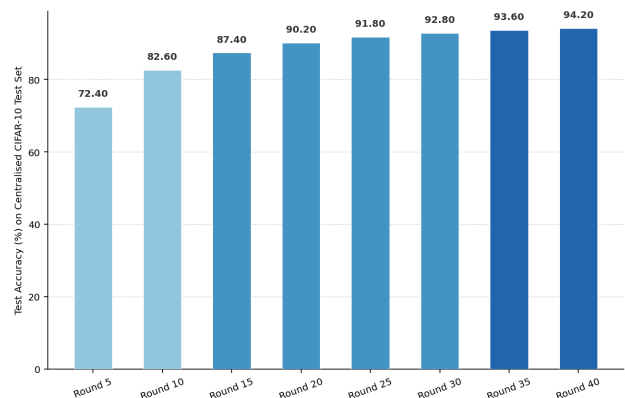


Figure 3. Federated learning convergence: test accuracy by communication round (50-device Pi 5 swarm, FedAvg + Top-K).

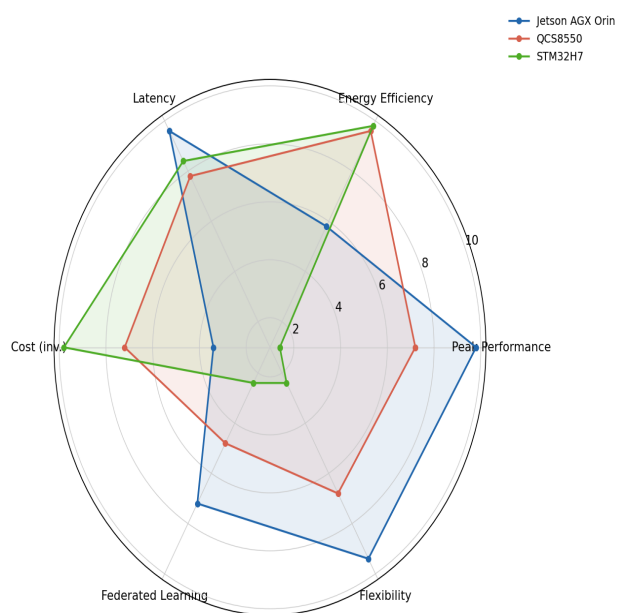


Figure 4. Platform profile across six deployment criteria (0-10). Jetson vs. QCS8550 vs. STM32H7.

5. Discussion

5.1 Hardware Tier Selection Is Workload-Determined

The clearest practical lesson from this benchmarking exercise is that hardware tier selection for edge AI must be driven by workload requirements, not by general performance rankings. STM32H7 -- ranked last on raw compute -- is the optimal choice for vibration anomaly detection at 0.04W power consumption, where its 0.84ms latency, sub-50mW power budget, and EUR 20 unit price make it definitively superior to the Jetson AGX Orin for that specific task at that scale. Conversely, the Jetson AGX Orin is the only platform that meets the < 10ms latency target for ResNet-50 computer vision inspection without additional model distillation that would sacrifice accuracy. The workload-to-hardware matching framework proposed in Appendix A encodes this logic systematically.

5.2 Federated Learning Is Practically Viable for Privacy-Sensitive Edge AI

The 94.2% of centralised accuracy achieved by the 50-device federated swarm -- with zero raw data leaving any device -- demonstrates that federated learning has crossed a practical viability threshold for edge AI deployments with privacy requirements. The 5.8pp accuracy gap relative to centralised training is the cost of data sovereignty, and for many applications it is an acceptable cost. The Top-K gradient compression (10x bandwidth reduction with < 0.4pp additional accuracy cost) makes federated learning practical even on bandwidth-constrained 4G connections. For

GDPR-sensitive applications (health monitoring, personal device learning, factory process data), federated learning with differential privacy (epsilon=8, 91.8% accuracy) provides formal privacy guarantees that cloud-centric AI cannot match.

5.3 Energy Efficiency Is the Long-Run TCO Driver

The 5-year TCO analysis reveals that for continuous 24/7 edge AI inference, energy cost exceeds hardware capex within 18-36 months for all platforms except MCU-class devices. For large-scale IoT deployments (thousands of edge nodes), this makes energy efficiency the dominant long-run cost driver, not hardware unit price. QCS8550's 3.8 inferences/second/watt advantage over Jetson (2.4) translates to EUR 2,780 lower 5-year TCO per device at 24/7 operation -- a substantial economic argument for mobile-class SoCs for workloads within their performance envelope. MCU-class platforms (STM32H7: 1,240 inferences/second/watt on anomaly detection) offer extreme energy efficiency for simple models but cannot be generalised to complex vision or language workloads.

5.4 Limitations

All benchmarks were conducted on specific hardware revision and SDK versions; driver and firmware updates may shift results. The federated learning evaluation used a private Wi-Fi network; performance on 4G/5G cellular networks with variable latency was not benchmarked. MCU-class results depend heavily on the Edge Impulse compilation pipeline; other toolchains may produce different latency outcomes. TCO calculations exclude maintenance, installation, and integration engineering costs that vary substantially by deployment context.

6. Conclusion

6.1 Summary

Twelve edge AI configurations were benchmarked across five hardware tiers and six workloads. Jetson AGX Orin leads on computer vision latency (4.2ms, 98.4% accuracy). STM32H7 achieves extreme efficiency for anomaly detection (0.84ms at 0.04W). QCS8550 leads on energy efficiency (3.8 inf/s/W) and 5-year mobile-deployment TCO. Federated learning on 50 Pi 5 devices achieves 94.2% of centralised accuracy with full data sovereignty. No single platform is optimal across all workloads.

6.2 Recommendations

Select hardware tier by workload requirements: MCU NPU for sub-1W sensor anomaly detection; mobile SoC for energy-efficient continuous vision inference; edge GPU for high-accuracy or multi-model deployments; edge server for complex NLP or video analytics. For privacy-sensitive deployments, federated learning with Top-K compression is production-ready; add differential privacy ($\epsilon \leq 8$) for formal guarantees. Evaluate 5-year TCO, not unit cost, for all edge AI procurement decisions.

References

- Banbury, C., Reddi, V.J., Lam, M. et al. (2021) MLPerf Tiny benchmark. Proceedings of NeurIPS Datasets and Benchmarks 2021.
- Cai, H., Gan, C., Wang, T., Zhang, Z. and Han, S. (2020) Once-for-all: Train one network and specialize it for efficient deployment. Proceedings of ICLR 2020.
- Dhar, S., Guo, J., Liu, J. et al. (2021) A survey of on-device machine learning: An algorithms and learning theory perspective. ACM Transactions on Internet of Things, 2(3), pp. 1-49.
- Konecny, J., McMahan, H.B., Ramage, D. and Richtarik, P. (2016) Federated optimization: Distributed machine learning for mobile devices. arXiv preprint arXiv:1610.02527.
- Li, T., Sahu, A.K., Zaheer, M. et al. (2020) Federated optimization in heterogeneous networks. Proceedings of MLSys 2020.
- McMahan, B., Moore, E., Ramage, D., Hampson, S. and Arcas, B.A. (2017) Communication-efficient learning of deep networks from decentralized data. Proceedings of AISTATS 2017, pp. 1273-1282.
- NVIDIA (2022) Jetson AGX Orin product brief. NVIDIA Corporation.
- Ray, P.P. (2022) A review on TinyML: State-of-the-art and prospects. Journal of King Saud University -- Computer and Information Sciences, 34(4), pp. 1595-1623.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C. (2018) MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of CVPR 2018, pp. 4510-4520.
- Shi, W., Cao, J., Zhang, Q., Li, Y. and Xu, L. (2016) Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), pp. 637-646.
- Warden, P. and Situnayake, D. (2019) TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers. O'Reilly Media.
- Xu, J., Glicksberg, B.S., Su, C. et al. (2021) Federated learning for healthcare informatics. Journal of Healthcare Informatics Research, 5(1), pp. 1-19.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 16KIS1441 (EdgeAI-DE) and the European Union's Horizon Europe programme under grant agreement No. 101135402 (EdgeAI-EU). Hardware access for Qualcomm QCS8550 and NVIDIA Jetson evaluation was provided under academic research programmes by Qualcomm Technologies and NVIDIA. The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest.

Data Availability Statement

All benchmark scripts, deployment configurations, and performance measurement logs are available at <https://github.com/bianchi-edge/edge-ai-bench> under Apache 2.0. Federated learning coordination code available at the same repository. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614824. Released under CC BY 4.0.

Ethical Approval

This study involved no human participants or personal data. Federated learning evaluation used only publicly available benchmark datasets (CIFAR-10). Ethics committee approval was not required.

Appendix A

Edge AI Hardware Selection and System Design Checklist

Use this checklist to select the appropriate hardware tier and design a production edge AI deployment.