RESCON TECHNOLOGIES

BareML™ Anomaly Detection Benchmark

July 2025

Abstract

ResCon Technologies was founded in 2020 with a mission to revolutionize the Edge AI space by creating the most computationally efficient machine learning algorithms possible. To that end, we have built BareML™, a software toolkit that enables the creation of incredibly fast, low power, and low latency machine learning models tailored for deployment to the smallest of embedded computers. BareML™ models shine brightest when speed and efficiency are paramount and training data is limited.

This white paper describes in detail our methodology and results for evaluating a BareML™ model using the Anomaly Detection use-case set forth in the MLPerf Tiny benchmarking framework. We find that our model's accuracy **exceeds the task's quality target** while evaluating over **88,000x faster** and using **41,000x less power** than the MLPerf baseline model.

Through market research we have identified several entities that have also beaten the original benchmark by significant margins. We still find a 220x improvement in evaluation speed and a 310x improvement in energy consumption over the most efficient of these models.

Finally, though not specifically measured in the MLPerf Tiny benchmarking framework, BareML™ models achieve state-of-the-art accuracy while displaying **ultra-fast training times on minimum data.** The results presented here arise from an algorithm training time of mere milliseconds on standard laptop hardware. This implies that BareML™ allows the entire machine learning model workflow to be accomplished locally on non-specialized hardware if desired—even directly on the target microcontroller.

Introduction

ResCon Technologies takes a fundamentally different approach to machine learning (ML). We believe that true innovation lies in superior algorithms, not brute-force computation. BareML™ puts this philosophy to work to create models that provide both high-quality results and utmost efficiency.

While the promises of ubiquitous AI, smart devices, and an interconnected world are immense, the reality is full of bottlenecks and critical challenges including hardware, bandwidth, and municipal power grid limitations. The continuing evolution of AI requires the deployment of effective and persistent ML algorithms to edge devices. But how do you run an algorithm designed for a data center on a tiny embedded microprocessor?

The conventional industry approach tackles this challenge by attempting to shrink large, power-hungry ML models, originally designed for cloud deployments, until they fit onto edge hardware. This "trim-to-fit" strategy, however, is fundamentally flawed. It starts with an inherently inefficient algorithm design and leads to models that, while functional, remain vastly inefficient and energy intensive. Engineers are forced into a compromise between accuracy and computational complexity.

In contrast, ResCon contends that the optimal approach is not to scale down an already bloated paradigm, but to **build from the ground up with algorithmic efficiency as the core principle**, creating models that are natively suited for the edge.

Deploying ML to edge devices is certainly not without its own challenges. Memory is limited, power availability is miniscule, and maintaining low latency becomes difficult. To help foster innovation in this space, MLCommons released MLPerf Tiny, a benchmarking suite designed for “apples-to-apples”

comparisons of Edge AI models and hardware.[1] This provides a standardized and agreed-upon methodology of evaluating a host of ML approaches for Edge AI and IoT solutions, enabling both engineers and stakeholders to better gauge the effectiveness of their solutions.

This white paper explains our results for the Anomaly Detection (AD) task within MLCommon's benchmarking suite, built on the ToyADMOS (Anomaly Detection in Machine Operating Sounds) dataset. The AD task provides insights into how models perform when faced with audio-based time series data, essential for predictive maintenance of industrial machinery.

[1] https://arxiv.org/abs/1911.02549

Methods

MLPerf Tiny is widely regarded as the standard public benchmarking suite for the evaluation of Edge ML tools and methods. It provides a set of rules that ensure each benchmarking result is a fair, transparent, and standardized representation of that specific implementation. This allows for many of the nuances involved in ML to be accounted for and helps developers objectively gauge the efficacy of their solutions. Results are important for customers who truly want the state-of-the-art, as it provides data that meaningfully translates to the real world.

The Anomaly Detection task specifically highlights effective solutions in industrial maintenance and real-time monitoring application spaces. The training dataset is a collection of "normal" (*i.e.*, non-anomalous) operating sounds of 4 types of toy cars. Each toy car class has 1000 samples of 10-second-long audio clips of nominal machine operation. The testing dataset is composed of both the normal sounds and anomalous sounds to which the model was not exposed during training. The model is evaluated on how well it can classify the anomalous sounds using an Area Under the Curve (AUC) metric to quantify classification accuracy, providing clear insight into how well a model is able to separate the incoming sound sample.

To ensure a meaningful benchmarking result, we strictly adhere to the rules outlined in the MLPerf Tiny paper. Our Device Under Test (DUT) is the Nucleo-L4R5ZI development board, which includes a 32-bit ARM Cortex M4 processor. This relatively basic reference board is chosen to emphasize how BareML™ models perform on affordable, low-complexity, and low-power hardware. Comparable reference model deployments are defined as those on the same hardware, with no access to GPU, specific AI

acceleration blocks, or other vector processors. This guarantees that all comparisons made are centered around cost efficiency, energy consumption, and latency, irrespective of external hardware implementations.

For energy consumption measurements, we use the prescribed X-NUCLEO-LPM01A board for power isolation and an Arduino Uno as a serial communication manager to ensure all measurements performed on the DUT are properly isolated from external interference.

Results

General

BareML™ Latency	BareML™ Energy Consumption	BareML™ Quality (AUC)
23 µs	1.99 µJ	0.86

Results are generated using V1.2 of the MLPerf Tiny benchmarking suite, the most current release as of July 2025.

Figure 1 shows BareML™ benchmarking metrics compared to the baseline and a variety of leading models.

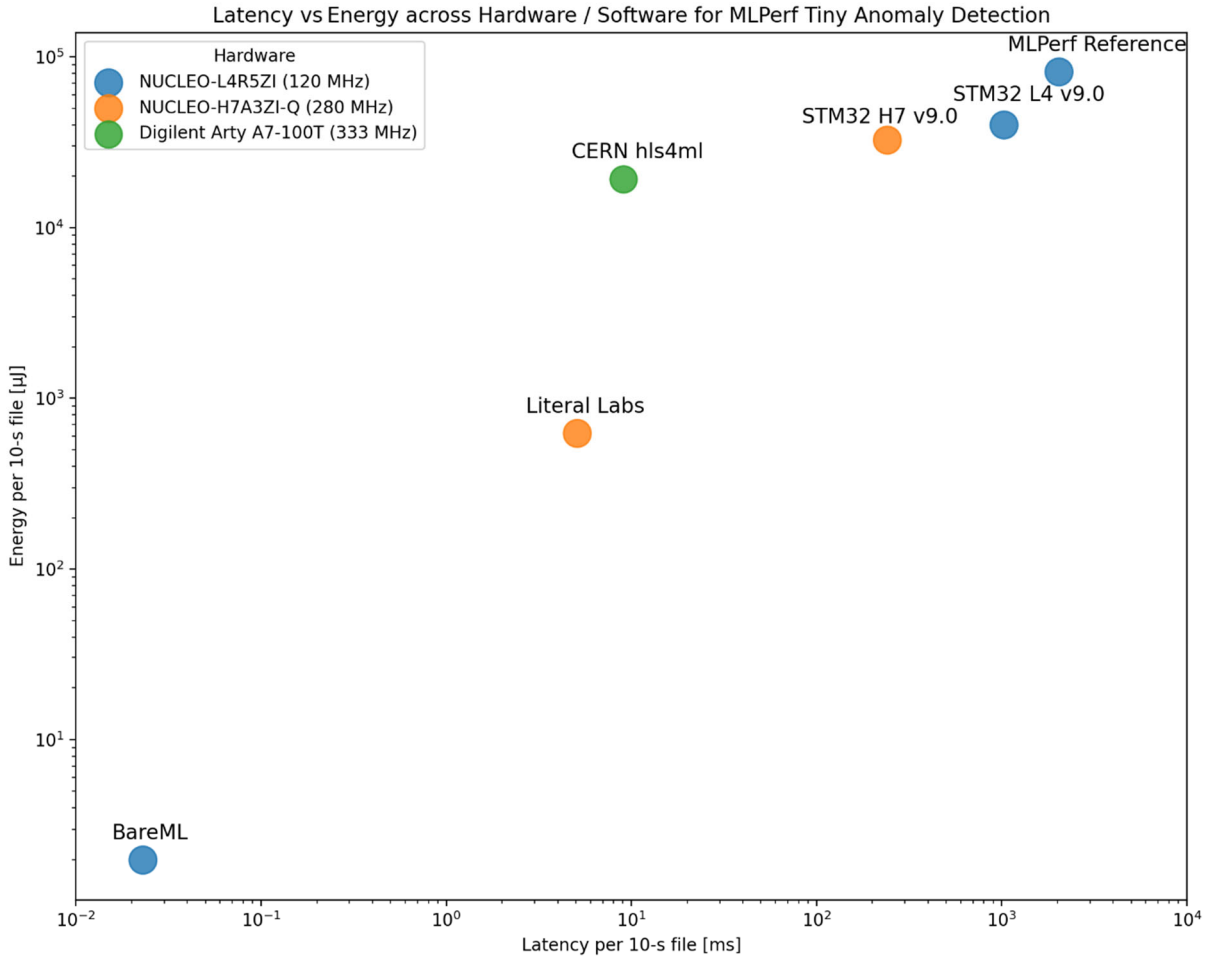

Figure 1. Results for a variety of model and hardware configurations performing the MLPerf Tiny AD benchmarking task. Metrics were obtained from ML Commons[2] except for Literal Labs.[3]

Comparing Models – Normalization

For the AD task, most conventional ML approaches employ a deep auto-encoder in single-stream mode, breaking down each full sample into *M* "windows," with each window defined as one inference. The BareML™ model, by contrast, *processes the entire sample simultaneously.* To ensure a fair comparison with window-based approaches, we normalize their reported metrics. Specifically, their latency and energy consumption values are multiplied by a factor of *M*, where *M* is the number of windows per sample employed in their benchmarking (typically, *M* = 196). This scaling allows metrics to be directly compared on a per-sample basis.

Importantly, **BareML™ still demonstrates superior benchmarking performance** even without this normalization.

Comparing Models – Hardware Selection

We perform BareML™ benchmarking using the NUCLEO-L4R5ZI (120 MHz) board, the same chipset used in the original MLPerf Tiny AD task. An examination of the Edge AI landscape, however, reveals a variety of boards used for benchmarking, including both microcontroller (MCU) and Field Programmable Gate Array (FPGA) hardware. This muddies the waters somewhat for an exact comparison of approaches. Figure 1 shows results from both the MLPerf Tiny Results website[2] and Literal Labs,[3] using color coding to differentiate the hardware platforms employed in each models' benchmarking.

[2] https://mlcommons.org/2024/04/mlperf-tiny-v1-2-results/

[3] See https://www.literal-labs.ai/. Disclaimer: Use of Literal Labs' publicly available benchmarking information is for comparison purposes only. No relationship between ResCon Technologies and Literal Labs exists, nor is any endorsement (or disparagement) of their products implied.

When comparing results in detail, it is important to understand the key differences in hardware. The top-performing model on the MLPerf Results page is CERN hls4ml,[4] benchmarked using the Diligent Arty A7-100T FPGA-based chipset. An FPGA has the potential to be computationally faster because it can highly parallelize the algorithm and uses a faster clock. Figure 2 shows that BareML™ results are still orders-of-magnitude better, despite the FPGA board's nearly 3x increase in clock speed and ability to run code in parallel.

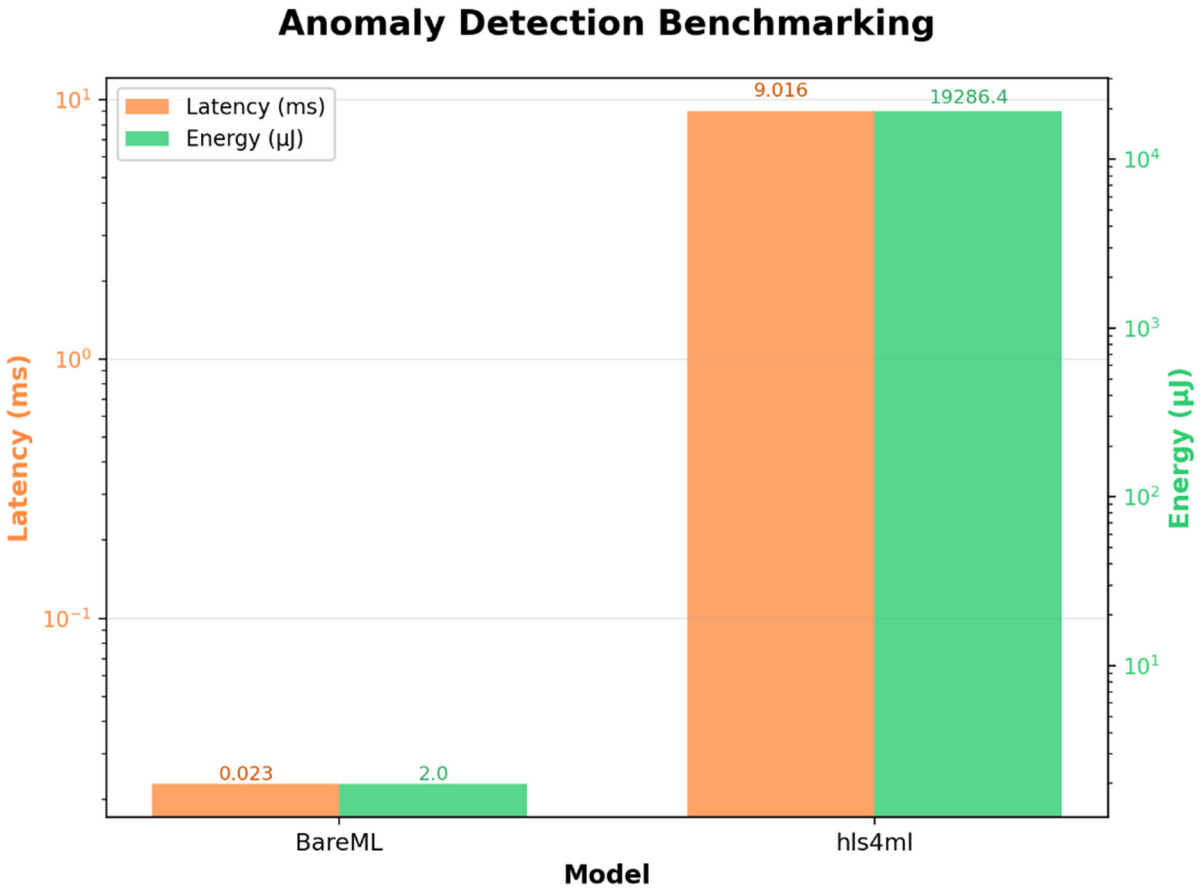

Figure 2. Direct comparison of BareML™ and CERN models, showing superior BareML™ metrics despite running on a slower and simpler hardware platform.

[4] https://arxiv.org/abs/2206.11791

Comparison to Other Ultra-Efficient Models

Market research shows that several companies have used alternative ML algorithms to generate impressive benchmarking results for the AD task. One of the most notable is Literal Labs, so we show a direct comparison between their reported results and BareML™. We assume that Literal Labs uses the same standard "windowed" approach to generate their metrics, so we normalize their results by multiplying by a factor of M = 196. With normalization, BareML™ shows a 220x improvement in latency and a 310x reduction in energy consumption over this leading approach (Figure 3).

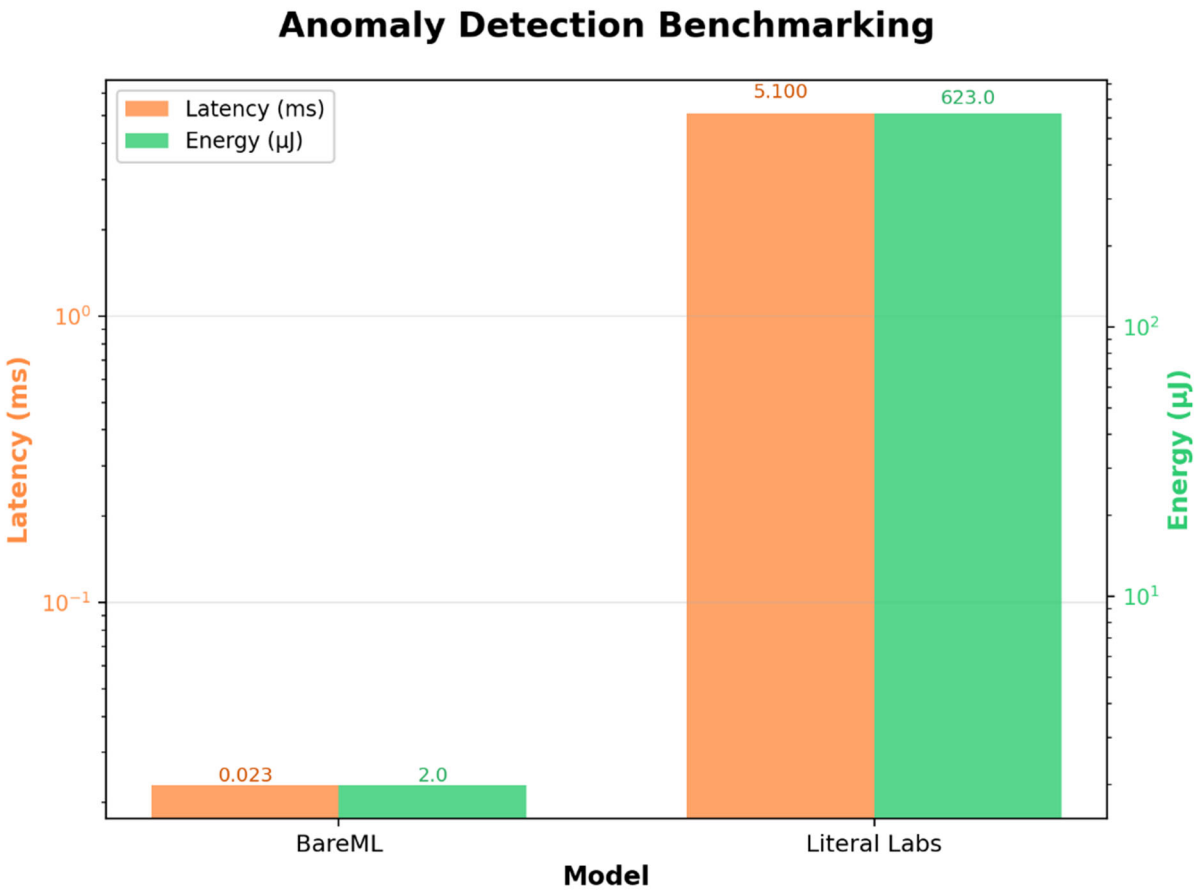

Figure 3. Comparison of BareML™ and an example of a leading ultra-efficient Edge AI model from Literal Labs.

Conclusion

The advantages of ResCon Technologies' BareML™ model creation toolkit and our efficiency-first approach to machine learning are clear. When benchmarked using the MLPerf Tiny suite, BareML™ does not merely show an incremental improvement over the state-of-the-art—it establishes a new standard of performance. The results confirm that our philosophy of prioritizing intelligent algorithmic design over brute-force computation yields solutions that are orders-of-magnitude superior for edge applications.

BareML™ delivers a 220x improvement in latency and a 310x reduction in energy consumption compared to the leading published results. These significant gains are achieved without compromising quality, as our model comfortably surpasses the benchmark's accuracy requirements. Furthermore, with the unique ability to be trained in milliseconds, our solution is not only powerful but also exceptionally practical, enabling local workflows and on-the-fly adaptation in dynamic environments.

The implications of these findings are game-changing. By moving beyond the limited "trim-to-fit" paradigm, we eliminate the perceived trade-off between performance and efficiency. BareML™ makes the deployment of robust, real-time AI on low-cost, battery-powered hardware not just possible, but practical and scalable. This unlocks new capabilities for a wide range of industries, particularly in the wearables and IoT spaces.

We invite you to explore how our innovative solutions can transform your Edge AI capabilities. To learn more about our technology or to discuss how our models can be tailored to your specific needs, please visit www.flyrescon.com and www.bareml.com, and contact us directly at support@flyrescon.com.