Thermal Characterization of AI Applications on AI Accelerators-equipped Microcontrollers

SiYoung Jang
Nokia Bell Labs
Cambridge, UK
siyoung.jang@nokia-bell-labs.com

Fahim Kawsar
Nokia Bell Labs and University of Glasgow
Cambridge and Glasgow, UK
fahim.kawsar@nokia-bell-labs.com

Chulhong Min
Nokia Bell Labs
Cambridge, UK
chulhong.min@nokia-bell-labs.com

ABSTRACT

We investigate the thermal characteristics of tiny ML devices with AI accelerators, to understand the thermal impact on wearable devices. We identify the unique characteristics of temperature measurement, temporal and spatial temperature variation, define two novel temperature metrics, saturation temperature and rate, and report these metrics for each compute component. For the analysis, we conduct various benchmarks and evaluate the thermal profiles of different AI applications and tasks. The results and insights lay an empirical foundation for the development of heat-safe wearable hardware.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; • **Hardware** → *Thermal issues*; • **General and reference** → Measurement.

KEYWORDS

AI accelerator, TinyML, Thermal measurement

ACM Reference Format:
SiYoung Jang, Fahim Kawsar, and Chulhong Min. 2024. Thermal Characterization of AI Applications on AI Accelerators-equipped Microcontrollers. In *Workshop on Body-Centric Computing Systems (BodySys '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3662009.3662020

BodySys '24, June 3–7, 2024, Minato-ku, Tokyo, Japan

ACM ISBN 979-8-4007-0666-0/24/06
https://doi.org/10.1145/3662009.3662020

1 INTRODUCTION

The development of microcontroller (MCU)-based tinyML devices such as Analog MAX78000 Feather [10], and Google Coral Micro [5] is quickly reshaping how and where artificial intelligence (AI) can be applied, making it more ubiquitous and integrated into our daily lives. Compact AI accelerators within such devices, like the MAX78000 measuring just 8mm × 8mm, enhance the efficiency of AI computations in devices. This allows compact wearable devices [8, 12] to run AI continuously and offer a variety of situational services.

Despite significant research efforts to enhance processing and energy efficiency in these tinyML devices, the critical aspect of thermal efficiency and its management has received less attention in the research domain. As overheating can lead to device failure and reduced lifespan [2], temperature throttling, which limits power once hardware chips reach a thermal limit, is commonly implemented to mitigate these risks. However, the thermal characteristics of AI applications and the impact of temperature throttling on service of quality for system-level heat management have not been actively explored yet. The challenge is further complicated by the nature of wearable devices, which are directly attached to the body. Studies [3, 13] have indicated that even a moderate temperature around 40C can cause skin injuries.

We explore the thermal characteristics of tinyML devices, aiming to understand how these compact, efficient AI accelerators generate heat while serving AI applications on wearable devices. While thermal profiling and management have been studied for higher-end devices like NVIDIA Jetson boards [1] and smartphones [9], a detailed analysis for MCU, the primary platform for wearable devices, is still lacking. Our study focuses on two AI accelerator-equipped MCU platforms, Google Coral Micro [5] and Analog MAX78000 featherboard [10] and conducts detailed benchmarks at both macro and micro levels. For the macro benchmark, we develop three AI applications and characterize their thermal footprints to understand the insights of thermal trend. Then, we perform the micro benchmark by focusing on primitive tasks to break down the thermal footprint of AI applications.

The key insights from our experiments are threefold: First, it is important to consider spatial and temporal factors, as heat distribution varies across the system and over time, depending on the applications. Second, there is a nonlinear correlation between power consumption and heat generation. Third, prioritizing CPU temperature management is crucial.

2 BENCHMARK SETUP

Benchmarking thermal characteristics of AI accelerators involves a comprehensive evaluation of their performance under various operational conditions. For this, we evaluate these devices across a range of settings while subjecting them to realistic workload scenarios.

2.1 Hardware Platforms

Coral Micro [5]: The Google's Coral Micro (see Figure 1 (left)) is an AI accelerator-equipped tinyML device designed to bring Google's powerful edge AI capabilities into compact and energy-efficient form factors. This platform is optimized for machine learning (ML) inference, making it ideal for portable and wearable technology. It features the Edge TPU coprocessor, a tailor-made hardware accelerator for TensorFlow Lite models, which is capable of performing up to 4 TOPS while maintaining a small footprint. The Edge TPU is configured to operate at 500 Mhz for the experiment. The core of the device is powered by Arm Cortex-M7 and Cortex-M4, ensuring a balanced performance for general computing tasks alongside AI workloads. Its maximum input supply voltage is 5 V.

MAX78000FTHR [10]: The Analog MAX78000FTHR (see Figure 1 (right)) is a platform engineered to ML inference at ultra-low power. It combines dual-core MCUs (Arm Coretex-M4 and RISC-V) and a specialized convolution neural network (CNN) accelerator, achieving up to 100× reduction in both latency and power consumption compared to running AI only on MCUs. The board includes MAX20303 Power Management Integrated Circuits (PMICs) that supports the input voltage of 3.7 V.

For each device setup, we used a Lithium ion polymer battery with 3.7 V and 500 mA to imitate real world wearable devices. Additionally, to observe the thermal impact of higher voltages, a battery pack with 5 V and 10000 mA battery is also used for on Coral Micro (Section 4.3).

2.2 Measurement Setup

2.2.1 Temperature measurement methodology. To measure the temperature of our target platforms, we used two primary methods: utilizing an on-chip *temperature sensor* and employing an external *thermal camera*. This dual approach

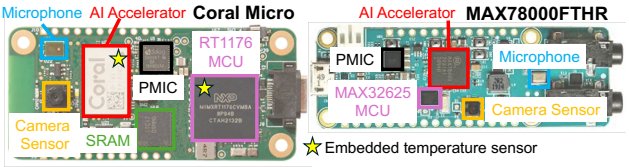


Figure 1: Compute element locations of Google Coral Micro [5] (left) and Analog MAX78000FTHR [10] (right)

to temperature measurement allows for a thorough understanding of the thermal dynamics affecting not only the core components but also peripheral ones, such as SRAM, camera sensors, and others. For our experiments, we kept the ambient room temperature at 25°C.

On-chip temperature sensor: Many embedded boards are equipped with built-in temperature sensors that provide direct and precise measurements of the chip's temperature. This allows for real-time monitoring of its thermal state at the system level.

Thermal camera: We measure the skin-contact point temperature using a Flir One Pro [7] thermal camera, which allows for a non-contact method to observe the spatial and temporal thermal changes across the device's surface.

2.2.2 Energy measurement. We also measure energy footprints using a Monsoon High Voltage Power Monitor [11] at a 5000Hz sampling rate to track the power consumption trends, setting the input voltage to match the battery's for benchmarking.

3 UNDERSTANDING THERMAL MEASUREMENTS

While most system resource footprints are presented as a single numerical value, such as 30% CPU usage or 15 mW power cost, presenting thermal footprints poses significant challenges due to the temporal and spatial variations in temperature. As heat generation and dissipation change over time, the same task can result in different amounts of heat depending on the previous temperature. Similarly, heat distribution is not uniform across the device's surface, as different electronic components generate varying amounts of heat.

3.1 Temporal Temperature Variations

We explain the characteristics of thermal footprints in comparison with energy footprints, which are regarded the most relevant to heat generation. Thermal footprints and energy costs, often considered proportional, actually differ due to distinct physical principles. Figure 2 shows a comparison of such characteristics during application execution. The energy footprint of a task is static and can be expressed as a single numerical value such as joules, because it is unaffected by temporal variables. Although the instant power within

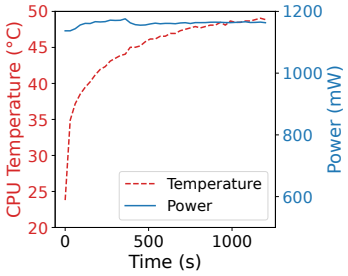
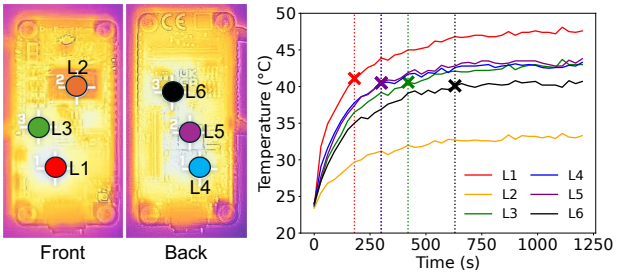


Figure 2: Heat and power consumption over time on Coral Micro.



(a) Thermal measurements on front & back side.

(b) Changes in temperature across various points.

Figure 3: Heat localization on Coral Micro.

a task execution can be dynamic, the energy consumed to perform a specific task remains constant regardless of the device's previous operational state.

In contrast, the thermal footprint is dynamic and significantly influenced by the device's previous thermal state, as dictated by the laws of *thermodynamics* [4]-—specifically, the first law, which relates to the conservation of energy, and the second, which addresses entropy increase in isolated systems. Thus, a device's heat output depends not only on the current task but also on residual heat from previous activities. Furthermore, due to *thermal inertia*-—a property of materials that quantifies the rate of heat absorption and release—-temperature changes within a device are gradual. Moreover, the temperature of a device does not increase indefinitely; it approaches a saturation point where the rate of heat generation equals the rate of heat loss, stabilizing the temperature.

Therefore, temperature reporting should be done in a way that reflects the nature of its temporal variation which involves using *thermal trends* to convey changes over time, rather than a single value. To this end, we define two thermal metrics: (a) the *saturation temperature* and (b) *saturation rate* by calculating the derivative of the temperature across measurement intervals.

Saturation temperature (T_s): First, we express the approach using the following equation: $\frac{\Delta T}{\Delta t} \leq \alpha$, where ΔT is the change in temperature between consecutive measurements, Δt is the time interval between these measurements and α is a predetermined threshold value. When the rate of change in temperature $\frac{\Delta T}{\Delta t}$ is less than or equal to α for β consecutive times, it indicates that the temperature change is minimal, and the component is considered to have reached its saturation point. For this work, we set the measurement interval, α, and β to 30 seconds, 0.02, and 3, respectively.

Saturation rate (R_s): We define the *saturation rate* by the following expression: $R_s = \frac{T_s - T_b}{t_s}$ where T_s, T_b refers to the saturation temperature and beginning temperature respectively. t_s refers to the amount of time till saturation in seconds.

3.2 Spatial Temperature Variations

Temperature variations across a device's surface demonstrate the non-uniform dynamics of heat distribution due to different hardware components generating varying heat levels during distinct tasks. Energy consumed by each hardware units also varies, but can be aggregated to reflect total energy consumption because energy has a single input source like battery which allows us to analyze its overall impact. However, understanding thermal patterns requires a spatial approach, especially in such devices that contact the skin, because temperatures at specific points can critically affect user's comfort and safety. Additionally, temperature at any point is affected by not only past thermal conditions but also the current operations of nearby components. Thus, in this paper, we will report the temperature measurements in three representative points: CPU, AI accelerator, and SRAM.

Figure 7 depicts the spatial nature of temperature variations of the Coral Micro. Figure 7a shows six locations of the thermal measurements and Figure 7b shows the change in temperatures over time at the corresponding locations. The results show that, the saturation temperature and rate differ significantly depending on the location. According to the reference [13], skin damage may occur at temperatures of 40°C. The marker **X** in Figure 7b represents the time it takes for each location to reach the threshold. All locations, except for L2, surpass 40°C. Specifically, L1 takes 180s, L3 takes 420s, L4 and L5 each take 300s, and L6 takes 630s.

4 THERMAL BENCHMARKS

4.1 Application Workloads

For the benchmark, we develop and experiment three applications that can run on AI accelerator-equipped wearable devices, each predominantly using different components.

- **Live scene analytics** uses computer vision and ML models on real-time video feeds from cameras to interpret and extract information continuously, warning users of

potential hazards. We utilize quantized and Coral Micro compatible version of SSD_MobileNetv2, and run it continuously.

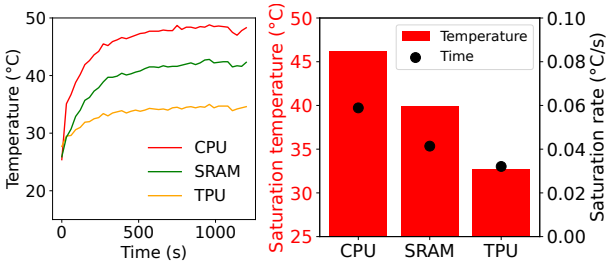
- **Life logger** records surrounding events for daily logging, using video feed and SSD_MobileNetv2 to analyze scenes similar to *live scene analytics* but at 0.2Hz since it does not require immediate responses.
- **Keyword spotting** detects specific keywords in real-time audio streams, allowing smart devices to respond to voice commands. It uses lighter Keyword spotting (KWS) model compared to those for object detection, resulting in lower computational demand and reduced AI accelerator usage. Audio signals are collected and KWS model inference is triggered every two seconds.

4.2 Thermal Comparison among Applications

4.2.1 Google Micro Coral. Figure 4a, 5a and 6a shows the temperature change over time of three locations—MCU, SRAM, and TPU— on the Coral Micro board with a battery pack of 3.7 V while running three different applications. The results indicate that the *Live scene analysis* application generates considerably more heat (up to 48°C) than the others, attributable to continuous image sensor readings, frequent memory I/O operations, and more ML inference. Although the same model was used, *Life logger* application produces a lower thermal impact than *Live scene analysis* due to lower ML inference cycles, providing intervals for passive cooling (up to 41°C). *Keyword spotting* application shows a trend similar to *Life logger*, as it includes two second interval between inferences to collect and process audio signals, which is less computational intensive reaching up to 42.6°C.

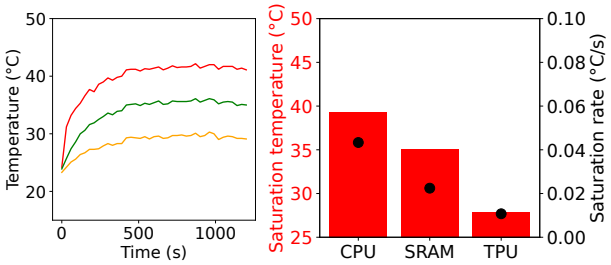
The three application's saturation temperature (T_s) and rate (R_s) are shown in Figure 4b, 5b and 6b. Despite the thermal footprints appearing similar in Figures 5a and 6a, they can be differentiated based on a combination of each component's T_s and R_s. Specifically, T_s for SRAM (33.5°C) and the R_s for the TPU (0.014°C/s) is lower in *Keyword spotting* compared to T_s for SRAM (35.1°C) and R_s for the TPU (0.022°C/s) in *Life logger*.

The results also show an interesting relationship between energy cost and produced heat. As expected, *Live scene analytics* consumes the highest power on average, 1035.4 mW and generates the highest temperature. The average power consumption of *Life logger* consumes around 633.3 mW, while *Keyword spotting* consumes 750.9 mW. While the difference in power consumption is significant (almost 18%), it's not straightforward to directly correlate this to the amount of heat generated. One reason is that the heat generated by each hardware component differs, depending on the type and intensity of the software task, even if the same amount



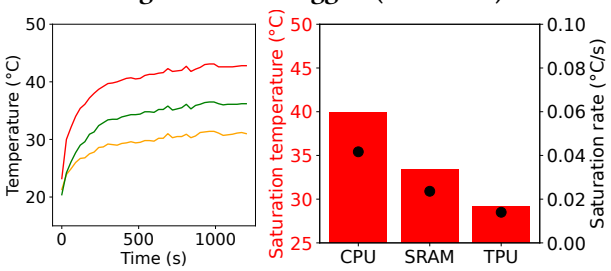
(a) Temperature. **(b) Saturation temperature.**

Figure 4: Live scene analytic. (1035.4mW)



(a) Temperature. **(b) Saturation temperature.**

Figure 5: Life Logger. (633.3mW)

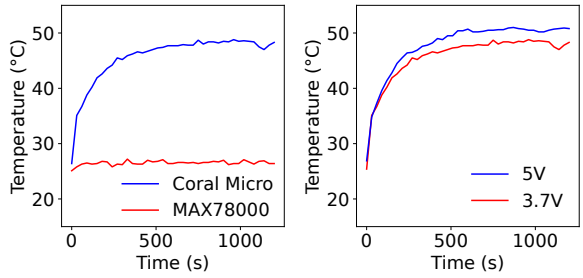


(a) Temperature. **(b) Saturation temperature.**

Figure 6: Keyword spotting. (750.9mW)

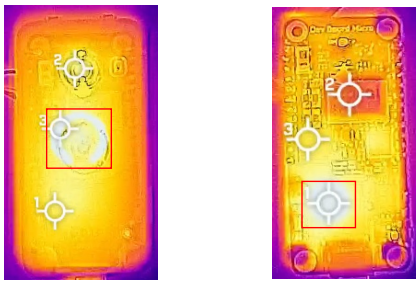
of total power is drawn. This also implies that heat profiles can not be simply estimated by the energy cost.

4.2.2 MAX78000FTHR. Figure 7a shows the temperature of MCU, which often shows the highest temperature on the surface, in comparison of MAX78000FTHR and Coral Micro while running *Live scene analytics*; both boards are powered on a battery pack of 3.7 V. The results show that, while a noticeable increase in change is observed on Google Coral Micro, the temperature for MAX78000FTHR board stays around the ambient temperature (25°C) over the full duration of the experiment. We conjecture that this is mainly due to the use of the different processor chip. As a main processor, the MAX78000FTHR uses an energy-efficient ARM Cortex-M4 core, while Coral Micro, on the other hand, uses a more powerful ARM Cortex-M7 processor. Due to the lower power profile of ARM Cortex-M4, the MAX78000FTHR tends to generate less heat, contributing to a smaller thermal footprint. This is also verified from the power cost. Even for the



(a) Ultra-low power tinyML device. (b) Input voltage.

Figure 7: Temperature analysis in different use case scenarios.



(a) With enclosure. (b) Without enclosure.

Figure 8: Heat distribution.

same input voltage, 3.7 V, the average power consumed by MAX78000FTHR and Coral Micro is 65.87 mW and 1035.4 mW respectively. As the noticeable heat changes are not observed in MAX78000FTHR even with the highest workload, we omit its footprint results in the next sections.

4.3 Impact of Input Board Voltage

The amount of voltage supplied to devices equipped with AI accelerators can affect the intensity of the heat these devices produce. Figure 7b illustrates the temperature of CPU over time comparing the use of 3.7 V and 5 V battery as input voltages for the Coral Micro. As expected, the results indicate that using a higher voltage results in approximately 3°C more heat. This is also clearly shown from the energy cost. Even for the same application, the average power was 1035.4 mW and 1082 mW when 3.7 V and 5 V were provided, respectively.

The CPU generates 5% more heat when the board is supplied with 5 V than at 3.7 V. This increase in heat can be attributed to several factors. One contributing factor is that when supplied with higher input voltage and the resistance is static, it is natural that the current increases dictated by the Ohm's law. As a result, the increased current leads to greater heat generation.

4.4 Impact of Enclosure

We conducted a case study comparing the temperature of Coral Micro with and without an enclosure. Note that the

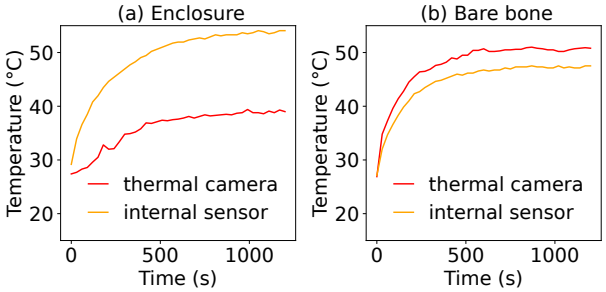


Figure 9: Temperature of CPU on Coral Micro board (a) with and (b) without enclosure.

results of this experiment is not applicable to all enclosures due to differences in size, material, and shape may impact different outcomes. For the enclosure of Coral Micro board, we have used the plastic case manufactured by Google [6] and put a battery pack as well underneath the board to simulate a real wearable device.

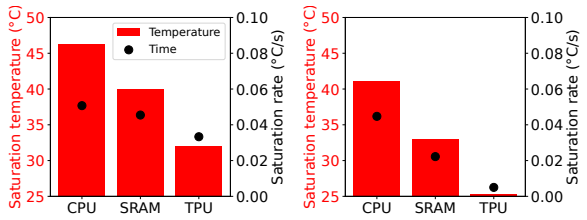
Figure 8a and Figure 8b show the heat distribution of *Live scene analytics* when operating at 3.7 V, with and without an enclosure, respectively. As presented earlier, when an enclosure was not used, heat hotspots are observed close to the hardware components that generate heat. Interestingly, with a casing, as depicted in Figure 8a, heat is spatially distributed inside while highest temperatures are focused near the outlets, highlighted in red box in each scenario.

We further compare temperatures measured from a thermal camera and an on-chip sensor. Figure 9 (a) shows the CPU temperature in a Coral Micro enclosure, revealing a significant disparity between sensor readings and thermal imaging due to heat trapped by the casing. The results indicate that the temperature builds up within the casing and eventually exceeds 50°C, which are record-high values in comparison to results obtained without the casing.

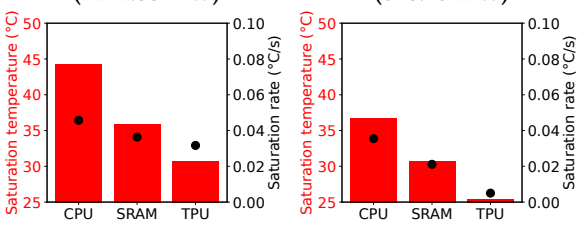
Figure 9 (b) illustrates the temperature changes for a Coral Micro bare-bone setup. Ideally, both measurements should align, but there is a 4°C discrepancy, possibly due to sensor placement, emissivity settings, reflective surfaces, calibration errors, and other environmental factors.

4.5 Micro Benchmark: Primitive Tasks

To further understand the different heat profiles of applications, we carry out an in-depth analysis of the thermal impact of primitive tasks. While continuously running these individual tasks alone in isolation does not reflect typical application behavior, it helps pinpoint the thermal effects of their execution. We elaborate four tasks which are commonly included in AI application pipelines and investigate their thermal characteristics when these tasks are executed continuously.



(a) ML model inference. (1144.55 mW) **(b) Memory I/O. (648.23 mW)**



(c) Image capture. (764.78 mW) **(d) Idle. (569.23 mW)**

Figure 10: T_s and R_s per task.

- **ML model inference task** executes model using an AI accelerator.In Coral Micro, this operation is called Invoke(). We utilise quantised and Coral Micro compatible version of SSD_MobileNetv2.
- **Memory I/O task** accesses and retrieves data from the flash memory and buffer in SRAM. LfsReadFile() operation reads and buffers 6.4MB of data continuously.
- **Sensor I/O task** captures sensor from the hardware and write to memory. GetFrame() is used to continuously capture 324×324 image in Coral Micro.
- **Idle task** refers to a background process which has no other active tasks or processes requiring CPU time.

Figure 10 shows the T_s and R_s for each task, with the idle task serving as a baseline. The results show important implications. First, as expected, the T_s and R_s significantly vary depending on the task because each task primarily utilizes a specific operation and a subset of components. For example, the T_s and R_s (25.3°C, 0.014°C/s) are significantly lower for the TPU component during the *Memory I/O task* compared to *ML model inference task* (32°C, 0.033°C/s). Second, regardless of the task type, CPU always reports the highest T_s and R_s since it is continuously used to control other peripherals. This remains true even for idle tasks, which report a T_s of 37.36°C without any tasks. This indicates the importance of focusing on CPU heat management in the design of heat-safe wearable hardware. Third, the AI accelerator efficiently manages temperatures during continuous *ML model inference task*, maintaining an increase of no more than 7°C, due to the edge TPU's specialized hardware design that supports high parallelism of MAC (multiply-accumulate) operations in AI accelerators which makes them highly efficient and hence generates less heat.

5 KEY TAKEAWAYS & CONCLUSION

We explored the thermal characteristics of AI accelerator-equipped MCUs. Our findings underscore two critical aspects of thermal management in devices. First, the location of heat-generating components significantly affects device temperatures, necessitating multiple values for accurate representation. Second, we emphasize the importance of considering the temporal aspect of temperature data, rather than relying solely on basic metrics. We introduced novel metrics that combines saturation temperature and rate per component and assess using both macro and micro benchmarks. This comprehensive insight is particularly crucial for wearable devices, where the compact size requires strategic placement of electronic components within a constrained space to ensure user comfort and device efficiency.

REFERENCES

[1] Amirhossein Ahmadi, Hazem A Abdelhafez, Karthik Pattabiraman, and Matei Ripeanu. 2023. EdgeEngine: A Thermal-Aware Optimization Framework for Edge Inference. In *2023 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 67–79.

[2] Hussam Amrouch, Seyed Borna Ehsani, Andreas Gerstlauer, and Jörg Henkel. 2019. On the efficiency of voltage overscaling under temperature and aging effects. *IEEE Trans. Comput.* 68, 11 (2019), 1647–1662.

[3] Children Wearable may cause skin burns [n. d.]. Wearable Children's Thermometer May Cause Skin Burns. https://www.mddionline.com/regulatory-quality/wearable-children-s-thermometer-may-cause-skin-burns. Accessed: 30 Mar. 2024.

[4] Rudolf Clausius. 1879. *The mechanical theory of heat.* Macmillan.

[5] Coral Micro [n. d.]. Google LCC. https://coral.ai/products/dev-board-micro/. Accessed: 30 Mar. 2024.

[6] Dev Board Micro Case [n. d.]. Google LCC. https://coral.ai/products/dev-board-micro-case/. Accessed: 30 Mar. 2024.

[7] Flir Pro One [n. d.]. Teledyne FLIR LLC. https://www.flir.co.uk/products/flir-one-pro/. Accessed: 30 Mar. 2024.

[8] Taesik Gong, Si Young Jang, Utku Günay Acer, Fahim Kawsar, and Chulhong Min. 2023. Collaborative inference via dynamic composition of tiny ai accelerators on mcus. *arXiv preprint arXiv:2401.08637* (2023).

[9] Seyeon Kim, Kyungmin Bin, Sangtae Ha, Kyunghan Lee, and Song Chong. 2021. ZTT: learning-based DVFS with zero thermal throttling for mobile devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 41–53.

[10] Max78000FTHR [n. d.]. Analog Devices. https://www.analog.com/en/design-center/evaluation-hardware-and-software/evaluation-boards-kits/max78000fthr.html. Accessed: 30 Mar. 2024.

[11] Monsoon High Voltage Power Monitor [n. d.]. Monsoon Solutions. https://www.msoon.com/high-voltage-power-monitor. Accessed: 30 Mar. 2024.

[12] Arthur Moss, Hyunjong Lee, Lei Xun, Chulhong Min, Fahim Kawsar, and Alessandro Montanari. 2022. Ultra-low power DNN accelerators for IoT: Resource characterization of the MAX78000. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 934–940.

[13] Smartwatch user skin burns [n. d.]. Samsung Galaxy Watch user claims wearable BURNED their wrist when left on while they were sleeping. https://www.dailymail.co.uk/sciencetech/article-11232191/Samsung-Galaxy-Watch-user-claims-wearable-BURNED-wrist-left-sleeping.html. Accessed: 30 Mar. 2024.