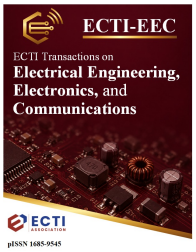
ECTI-EEC Transactions Pre-proof



THERMO-SAFE-RV: A Thermal-Aware Secure RISC-V Architecture for Edge-AI Nodes

Keshav Kumar, Mohit Kumar Srivastava, Vivek Kumar, Utkarsh Pandey, and Amandeep Kaur

DOI: http://doi.org/10.37936/ecti-eec.2027251.267397

Reference: ECTI-EEC 267397

Received date: 10 August 2026

Accepted date: 28 October 2026

This is a PDF of an article that has been improved after acceptance, including the inclusion of a cover page and metadata, as well as the optimization of readability. Before its publication in its final form, this version will undergo further copyediting, typesetting, and review. As a result, this version is no longer the Accepted Manuscript, but it is not yet the definitive **Version of Record**. We are providing this early version to provide the article with early visibility.

Published by ECTI Association Thailand

ECTI Transactions on Electrical Engineering, Electronics, and Communications, Vol. 25, No. 1, pp. 53–66, February 2027

https://doi.org/10.37936/ecti-eec.2027251.267397
ISSN(Online): 1685-9545

THERMO-SAFE-RV: A Thermal-Aware Secure RISC-V Architecture for Edge-AI Nodes

Keshav Kumar[1], Mohit Kumar Srivastava[2†], Vivek Kumar[3], Utkarsh Pandey[4] and Amandeep Kaur[5]

Abstract

Edge AI systems perform various tasks like perceiving, classifying, and controlling near the sensors while considering the constraints of temperature, resources, and security. RISC-V can be described as a flexible and open framework that could serve the needs of such systems, but traditional edge AI systems view the temperature control, security control, and scheduler for the accelerator as independent components, leading to delays created during the slowdown process, an increase in costs, and problems with unwarranted access to unauthorized programs from the attacker. The authors of this paper present THERMO-SAFE-RV, which integrates several elements of the technology, including dynamic voltage frequency scaling (DVFS), lightweight monitoring procedures, and boot assessment technology for the efficient edge AI. A composite risk–energy–thermal controller selects operating points from observed temperature, power, task criticality, and a runtime anomaly score and binds every admissible response to a cryptographically signed policy table, assuming a protected runtime enforcement path. A constrained optimization model is solved online by a table-driven search over a small candidate set. At 80.00% workload intensity, the reported architecture-level simulation reduces energy per inference from 8.70 mJ for Baseline RV to 5.50 mJ for THERMO-SAFE-RV and peak die temperature from 78.4°C to 64.9°C. Compared specifically with the Secure Monitor scheme, reported attack detection increases from 91.20% to 96.80%, and false alarms decrease from 7.80% to 3.10%. These are simulation estimates; calibration records, a consolidated hardware-overhead ledger, and implementation measurements are not available to establish absolute hardware performance.

Keywords: RISC-V, Edge AI, Embedded System Architecture, Dynamic Voltage and Frequency Scaling, Runtime Intrusion Detection, Thermal-Aware Architecture, Hardware/Software Co-Design, IoT Nodes.

I. Introduction

Edge-AI systems bring inference closer to the sensors, cameras, autonomous platforms, and industrial machines by bringing it to embedded nodes. This not only increases the network latency and limits access to private data but also moves time-consuming neural-network tasks into low-power processors, microcontrollers, and small accelerators with strict power and thermal constraints [1]–[6], [37], [38].

RISC-V's open ISA (instruction-set architecture), extensible modules, and customizability for accelerators have made it an appealing processor foundation. Complete system of control software, special CNN acceleration, memory protection, and runtime monitoring can be embedded in a single platform that is built using the RISC-V architecture [7]–[11]. There is a need for more flexibility and system-level co-design: a design that optimizes throughput may heat up too much, a design that secures security may introduce unnecessary latency, and a design that implements a simple DVFS may not detect abnormal behavior due to attacks.

Thermal reliability and security are closely related in resource-limited embedded systems. Switching activity is increased by abnormal code execution, peripheral abuse, and repeated memory transactions or adversarial input streams, which leads to thermal stress, and aggressive throttling at high temperature decreases the timing margin of real-time inference tasks. The coupling is a two-way one: an attacker who increases switching activity can cause a purely thermal governor to throttle, and a purely security- driven monitor who adds instrumentation overhead will increase the temperature that it is not monitoring. The above considerations should therefore be considered as coupled runtime signals that should be handled by a single decision rule in a practical architecture, not independent controllers that can make conflicting decisions.

Manuscript received August 10, 2026; revised on October 16, 2026; accepted October 28, 2026.
Recommended for publication by Associate Editor Krit Angkeaw.
†Corresponding Author: Mohit Kumar Srivastava (mohit.srivastava@psit.ac.in)
[1,2,3,4]Department of Electronics and Communication Engineering, Pranveer Singh Institute of Technology (PSIT), Kanpur, India.
[5]Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India.

A. Contributions

- A thermal-aware secure RISC-V edge-AI architecture in which DVFS, secure boot attestation, runtime intrusion monitoring, and CNN accelerator scheduling are coordinated by a single controller rather than by independent modules.
- A composite risk–energy–thermal objective (Section 5) that minimises energy under risk-dependent admissibility rules subject to latency and temperature constraints, with a table-driven online solution whose cost is bounded by the size of the operating-point set.
- A lightweight logistic anomaly score derived from memory-access bursts, control-flow events, peripheral activity and accelerator DMA behaviour, requiring only hardware event counters and no operating-system instrumentation, together with a slowly adapting threshold intended to limit false alarms during bursty workloads; feature sufficiency and resistance to adaptation poisoning remain unverified.
- A signed policy table as the trust anchor binding attestation, admissible frequency states, thermal thresholds, and intrusion responses, with protection against post-boot policy modification dependent on an enforceable privilege boundary.
- An architecture-level comparison of energy, latency, peak temperature, deadline satisfaction, and reported security metrics against Baseline RV, DVFS-only, Secure Monitor, and thermal-guard schemes. The existing ablation removes complete architectural blocks, and module-overhead values are preliminary estimates. Temperature-specific classification-accuracy claims are excluded because the documented controlled-operation results do not support them.

Section II reviews related work and states the research gap. Section III defines the node, threat and signal models; Section IV describes the architecture; Section V develops the optimization formulation and control algorithm; Sections VI and VII give the experimental setup and results; Section VIII discusses limitations and threats to validity; and Section IX concludes.

II. RELATED WORK AND RESEARCH GAP

A. RISC-V Edge-AI and Embedded Acceleration

Embedded AI systems combine programmable cores with neural accelerators to support local inference under tight power budgets [1]–[6], [38], and RISC-V enables domain-specific extensions and open co-design [7]–[11], with cryptographic instruction-set extensions now standardized for this class of node [40]. Such work nevertheless optimizes throughput and energy efficiency, treating runtime thermal and security interaction as secondary. A high-throughput accelerator is unsuitable if it produces unsafe thermal peaks or if its control path can be manipulated by compromised software.

B. Thermal-Aware Runtime Management

Thermal-aware design relies on DVFS, task migration, clock gating, and throttling [12]–[15], [39], and has been extended recently to neural workloads on NPU-equipped platforms [33] and combined with reliability and mixed-criticality objectives [34]–[36], which shows that multi-objective governors are practical on constrained hardware. None of these formulations includes a security-risk term, even though malicious workload patterns produce abnormal power and temperature behavior that a utilization-driven governor reads as legitimate demand.

C. Embedded Security and Runtime Intrusion Monitoring

The mechanisms of embedded security consist of memory protection, secure booting, control flow monitoring, attestation, and anomaly detection [16]–[22]. RISC-V trusted execution policies, such as Keystone [31] and the CoVE extensions [32], demonstrate that isolating and attesting primitives are needed, since actual threats to the hardware RISC-V configurations have proved the need for these primitives [41]. Lightweight monitoring systems are adequate for certain nodes that are unable to use any host-related solutions, but the monitoring-only infrastructures do not provide thermal safety, and systems not aware of the use of accelerators do not know about DMA, model-buffer, and peripheral-related attacks.

D. Hardware/Software Co-Design for Reliable Edge Systems

es, accelerators, and software [23]–[27], [37]. In edge AI, this encompasses neural layer mapping, precision determination, bandwidth management, and power state control [28]–[30]. The difficulty being faced is how to merge security threat with thermal state review into one decision-making loop without creating high complexity.

TABLE I
POSITIONING OF RELATED WORK AND THE RESPONSE OF THIS PAPER.

Research stream	Remaining limitation	Response in this paper
RISC-V edge AI [1]–[11], [38]	Security and thermal	Unified secure thermal edge-AI architecture (Section IV)
Thermal management [12]–[15], [33]–[36], [39]	No intrusion-risk term in the governor	Risk-aware thermal governor, Eqs. (7)–(13)
Embedded security [16]–[22], [31], [32], [41]	Energy and temperature impact of monitoring ignored	Joint risk, energy and temperature objective, Eq. (8)
HW/SW co-design [23]–[27], [37]	Limited runtime threat awareness	Policy-driven mapping with monitor feedback
Efficient low-power AI [28]–[30]	Reliability–security trade-off incomplete	Normal, guard and safe operating modes

E. Research Gap

There are four gaps present. The first gap is that security and thermal control are optimized in isolation, whereas thermal stress increases with attacks and abnormal workload patterns. The second gap is that DVFS mechanisms can neither differentiate between normal, high-load inference and abnormal behavior during runtime, and mechanisms based on monitoring are able to detect abnormal behavior but add processing time without coordinating frequency-setting processes. The third gap is that the temperature stability, detection ratio, false-alarm ratio, and accuracy are seldom measured at once for the same design point, making it difficult to properly evaluate trade-offs.

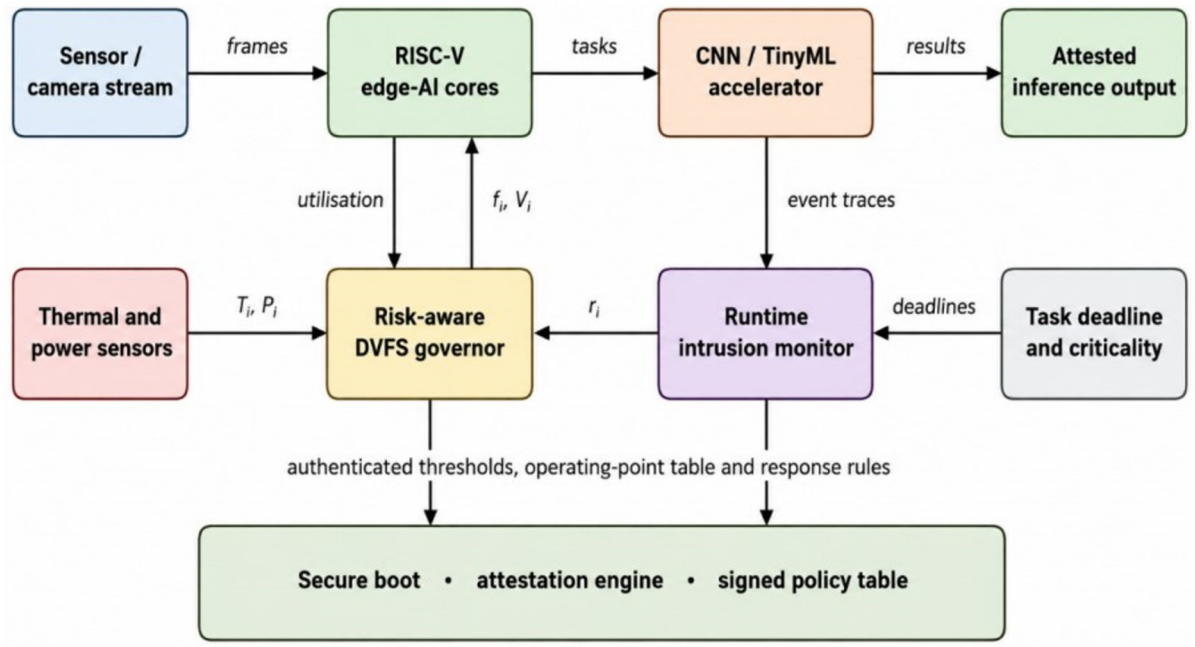


Fig. 1. Proposed THERMO-SAFE-RV Node Architecture and Runtime Signal Flow.

The fourth gap is that an approach that combines energy, risk, latency, and temperature in one constrained objective function and that is inexpensive to compute and allows evaluation at the level of dozens of milliseconds is not available yet. This paper investigates the architectural coordination problem within an analytical simulation; implementation cost, feature sufficiency, and workload transfer require further validation.

III. SYSTEM MODEL, THREAT MODEL AND RUNTIME SIGNALS

A. Edge-AI Node Model

The node here has a RISC-V core cluster, a lightweight CNN accelerator, a shared memory, an interface for sensors, a thermal sensor, a power sensor, a secure boot module, and a runtime monitor to facilitate its operations. Fig. 1 presents an image showing the relationships between the different parts of this architecture and the different runtime signals exchanged between the parts. In this case, the workload consists of a sequence of inference tasks, with each task having its deadline, model size, input resolution, and criticality level. The runtime controller monitors the temperature, power, utilization, and security issues and predicts the voltage-frequency operating point and accelerator mapping before the next control epoch occurs.

B. Threat Model

The trusted computing base consists of the boot ROM, attestation engine, policy table, and the hardware event counter. Everything above this base, including the application, the inference runtime, device drivers, and downloaded model, is untrusted. The adversary is assumed to have reached arbitrary code execution at the application or driver privilege level, such as by exploiting a model artifact, a memory safety defect in the sensor parsing path [19, 20], or a compromised firmware update channel. From this point the adversary can (i) cause abnormal memory access bursts causing timing degradation and increasing switching activity, (ii) program the DMA engine to read or write model buffers, (iii) switch peripherals to increase power consumption, and (iv) try to disable the monitor and/or select an operating point outside the certified Vt/Vf envelope.

There are two types of consequence that are in scope. The first is Thermal-reliability: If the die continues to exhibit abnormal activity, it will reach Tmax and become subject to reactive throttling, deadline misses, and aging. The second is confidentiality and integrity: Unauthorized access to the DMA to model buffers or inference outputs. Physical and invasive attacks do not apply (power and electromagnetic side channels, fault injection, and decapsulation), as they involve circuit-level rather than architecture-level attacks. Some of the most effective stealthy adversaries that adapt their activity to the benign event distribution are only partially addressed; this is revisited in Section VII.b.

Statistical anomaly detection and policy enforcement provide different protections. A high counter-based score can indicate abnormal resource use, but it does not establish that a DMA transfer was unauthorized or that confidentiality was preserved. Preventing policy bypass requires trusted control-register and counter configuration, protected policy storage, and an enforcement path inaccessible to the assumed application or driver attacker. These runtime protections are architectural requirements; boot-time signature verification alone does not validate them.

C. Workload, Power and Thermal Models

Let f_i denote the clock frequency selected at control epoch i, V_i the supply voltage, u_i the accelerator utilization, and T_i

the peak die temperature. The inference latency depends on the operating point and on how much of the layer graph is mapped to the accelerator,

$$L_i = \frac{L_0}{a_f f_i + \alpha_a u_i} + L_{mon} \tag{1}$$

where L_0 is the normalised workload size, αf and αa are the core and accelerator throughput coefficients, and L_{mon} is the fixed latency contributed by monitoring. The instantaneous power combines dynamic switching power, a voltage-dependent static component and the accelerator contribution,

$$P_i = C_{\text{eff}} V_i^2 f_i + P_{\text{stat}}(V_i) + P_{\text{acc}} u_i \tag{2}$$

so that the energy consumed by one inference is

$$E_i = P_i L_i + E_{\text{men}} + E_{\text{mon}} \tag{3}$$

with E_{mem} and E_{mon} the memory-traffic and monitoring energy per inference. Eqs. (2) and (3) make the coupling explicit: the operating point that reduces energy also reduces the power that drives the thermal state. Temperature evolution is modelled by a first-order discrete thermal equation, used here as a lumped controller-level approximation rather than a spatial or circuit-level thermal model [12], [14],

$$T_{i+1} = \rho T_i + (1-\rho) T_{\text{amb}} + \kappa P_i \tag{4}$$

where $0 < \rho < 1$ controls thermal memory, Tamb is the ambient temperature and κ maps power into thermal rise; a larger κ represents a tighter package or a lower heat-dissipation capability. Because Eq. (4) is a one-step predictor, the controller can evaluate the temperature that a candidate operating point would produce before committing to it, which is what allows guard mode to be entered before rather than after the limit is crossed.

The model treats each epoch as a lumped, approximately constant operating state. Eqs. (1)–(3) do not explicitly resolve memory contention, voltage-transition settling, spatial hotspots, or switching transients. Latency and power units must be consistent when energy is computed: power in watts multiplied by latency in milliseconds yields millijoules. Memory and monitor energy are additional contributions only if they are not already included in the fitted power coefficients; otherwise, they would be counted twice.

For the stated thermal coefficients ρ = 0.86 and κ = 4.2 °C W^{-1}, the steady-state rise implied by Eq. (4) is $\kappa P/(1\rho)$, corresponding to an effective resistance of 30.00 °C W^{-1}. If one recurrence step is interpreted as the 5 ms control epoch, the associated single-RC time constant is 5/ln(0.86) = 33.15 ms. These values are deductions from the recurrence, not package measurements. The initial temperature, ambient schedule and calibration residuals are not reported, so package-specific transient accuracy cannot be assessed.

D. Runtime Security-Risk Score

The monitor computes a normalised risk score from hardware event counters. The score is deliberately lightweight so that it can be evaluated every epoch on a small, embedded platform without operating-system support. Let the weighted evidence at epoch i be

$$z_i = \beta_1 m_i + \beta_2 c_i + \beta_3 p_i + \beta_4 d_i - \beta_0 \tag{5}$$

where m_i, c_i, p_i and d_i are the counts of abnormal memory bursts, control-flow alerts, peripheral anomalies and DMA access-pattern deviations observed in epoch i, normalized by the epoch length, and β_0 is a bias term. The risk score is the logistic transform of this evidence,

$$r_i = \sigma(z_i) = \frac{1}{1+e^{-z_i}} \in [0,1] \tag{6}$$

which requires one multiply–accumulate over four counters and a small lookup table for $\sigma(\cdot)$. This is an anticipated low-priced score; metrics such as its execution cycles, lookup-table size, and implementation burden have not been computed yet. An initial system indicates an off-line calibration on benign traces but does not provide details concerning coefficient values, event designations as well as the fitting process. Besides, a logistic transformation assures that the score value does not exceed one, thus being compliant with the required standard of event attack probability. The chosen four signals correspond to the interfaces affected by the stated threat model (i.e., memory traffic, execution control, peripheral activity and accelerator DMA). The grounds of the signal selection include their architectural coverage instead of relying on the method of determining the best panel of features.

Situating the four-feature score's adequacy is not supported through the current set of block experiments. Validation at the feature level requires a full score model and four leave-one-out counterparts built on the same dataset of stratified attacks and benign transcripts, with model masking that accounts only for all other features without retraining. To assess family attribution to attacks, measurements must be made regarding false alarms due to benign bursts and the sensitivity of thresholds/windows. Such testing is not presented among the current results.

TABLE II
RATIONALE AND COVERAGE LIMITS OF THE FOUR RUNTIME SIGNALS.

Signal	Threat-model link	Benign confounder	Coverage limit
Memory bursts	Memory-access flooding and resource exhaustion	CNN or sensor bursts	Low-rate or workload-matched abuse may evade detection
Control-flow alerts	Execution deviations following application or driver compromise	Legitimate execution variability depends on the alert definition	Allowed-path data misuse need not produce an alert
Peripheral anomalies	Device toggling and peripheral-power abuse	Legitimate alarms and acquisition bursts	Peripheral-specific attacks were not separately evaluated
DMA deviations	Access involving model buffers or inference outputs	Legitimate layer and model transfers	Traffic counts alone do not establish access authorization

IV. PROPOSED THERMO-SAFE-RV ARCHITECTURE

A. Architectural Overview

THERMO-SAFE-RV inserts a controller between the edge-AI runtime and the hardware resources. The controller does not replace the RISC-V execution stack; it observes system signals, classifies the operating state and selects a safe operating mode. The architecture has five functional blocks, shown in Fig. 1: secure boot and attestation, performance and thermal sensing, anomaly scoring, risk-aware DVFS, and accelerator scheduling. Their runtime interaction is executed once per control epoch and is stated as the intended controller procedure in Algorithm 1.

B. Secure Boot and the Signed Policy Table

At boot, the firmware image, the accelerator configuration, and the policy table are authenticated against a key held in the trust anchor. The policy table stores the allowable frequency and voltage states, the thermal thresholds T_g and $T_{\max}$, the risk thresholds τ_1 and τ_2, the anomaly weights and the response rules for each task-criticality level. Authentication establishes the integrity of the table when it is verified. Preventing post-boot modification or monitor bypass additionally requires write protection and privileged enforcement that remain inaccessible to untrusted application and driver code. A modified table can be rejected upon verification; continuous detection of every runtime modification is not established by the architecture-level simulation. Recovery mode is the intended response to verification failure. This mirrors the trust-anchor pattern used by open RISC-V trusted-execution frameworks [31], [32], and reuses the same measurement primitives [16], [18].

C. Runtime Intrusion Monitor

The monitor collects event counters from the memory, control-flow, accelerator, and peripheral interfaces and evaluates Eqs. (5) and (6). When the score exceeds a threshold, the controller raises the protection level, restricts DMA access to the model buffers, and shifts the system towards a safe inference mode. The mode is a joint function of risk and temperature rather than of either signal alone,

$$\begin{aligned} s_i &= \text{safe}, && \text{if } r_i \geq \tau_2 \text{ or } T_i \geq T_{\max}, \\ s_i &= \text{normal}, && \text{else if } r_i < \tau_i \text{ and } T_i < T_g, \\ s_i &= \text{guard}, && \text{otherwise}. \end{aligned} \tag{7}$$

If the high-risk threshold is reached, or the hard temperature limit is reached, then safe mode takes priority. If the risk is not low and the temperature is not below the guard, then they go into normal mode, otherwise they go into guard mode. τ_i is the adaptive low-risk threshold (initialized to τ_1) and τ_2 is the fixed high-risk threshold. This is an ordered definition where there is no overlap between the original guard disjunction and the high-risk safe condition. The previously reported averages are not sufficient to prove boundary behavior; it has to be explicitly validated.

D. Risk-Aware DVFS and Accelerator Scheduling

The governor chooses the lowest-energy operating point that satisfies the current deadline while avoiding thermal and security violations. When risk is low, the accelerator is used aggressively to reduce both latency and energy. When risk is high, memory access is restricted and a more conservative operating point is selected even if the latency overhead increases moderately. Formally, the governor minimizes energy augmented by risk and thermal penalties over the admissible triples (f, V, u),

$$\begin{aligned} (f_i^*, V_i^*, u_i^*) = \arg\min_{(f,V,u)} \\ [E_i + \lambda_r r_i + \lambda_T \max(0, T_{i+1} - T_g) \end{aligned} \tag{8}$$

The weights λr and λT convert risk and thermal excess into energy-equivalent units and are fixed at design time in the policy table. Eq. (8) is the design objective; Section 5 turns it into the composite runtime cost that the controller evaluates online, subject to explicit latency and temperature constraints.

The observed r_i is the same for every candidate in one epoch. Consequently, the additive $\lambda r\ r_i$ term does not change the ranking of candidates within the same feasible set. Risk affects selection through mode-dependent admissibility and protective response rules. A candidate-dependent estimate of future risk would be required for the risk penalty itself to influence within-set ranking; that extension is not evaluated here.

V. RUNTIME CONTROL FORMULATION AND ALGORITHM

A. Constrained Problem Statement

Over a horizon of N control epochs, the controller solves

$$\begin{aligned} \min_{(f_i,V_i,u_i,s_i)} \sum_{i=1}^{N} C_i, \\ C_i = E_i + \lambda_r r_i + \lambda_T \max(0, T_i - T_{\text{safe}}) \end{aligned} \tag{9}$$

subject to, for every epoch i,

$$\begin{aligned} L_i \leq L_i^{\max}, \quad T_i \leq T_{\max}, \quad f_i \in F, \\ V_i \in V, \quad u_i \in U, \quad s_i \in \{\text{normal}, \text{guard}, \text{safe}\} \end{aligned} \tag{10}$$

The sets F, V and U are the discrete frequency, voltage ,and accelerator-utilization levels admitted by the signed policy table. Because these sets are small and authenticated, the search space is bounded, while its trustworthiness depends on the runtime enforcement assumptions in Section IV.b, which is what makes a per-epoch exhaustive evaluation affordable.

Eq. (9) states a horizon-level objective; Algorithm 1 implements a per-epoch greedy search using a one-step thermal predictor, rather than a demonstrated globally optimal horizon solver. The policy-dependent target T_{safe} and energy-equivalent penalty coefficients are not specified numerically in the available configuration, which limits reproducibility of this formulation.

B. Composite Runtime Cost and Operating-Point Selection

A single metric is insufficient for an embedded controller, so the per-epoch decision uses a composite cost containing normalized energy, normalized latency, risk and thermal margin,

$$J_i = w_E \frac{E_i}{E_{\text{ref}}} + w_L \frac{L_i}{L_i^{\max}} + w_R r_i + w_T \max(0, T_{i+1} - T_g) \tag{11}$$

E_{ref} is the energy of the baseline operating point. The first three terms are dimensionless. Because the temperature excess in Eq. (11) is expressed in degrees Celsius without a normalizing temperature scale, w_T must carry inverse-temperature units for J_i to be dimensionless. The reported numerical settings 0.35, 0.25, 0.25 and 0.15 therefore cannot be described as a dimensionless convex combination merely because their numerical values sum to one. This revision retains the original cost expression and its numerical settings, and corrects the interpretation rather than introducing an untested normalization. The units and calibration convention must be documented before implementation.

The controller then evaluates the candidate voltage–frequency states and accelerator utilization levels that remain feasible under the real-time and thermal constraints,

$$\Omega_i = \{(f, V, u) \in F \times V \times U : (f, V) \in \Pi_{FV}, L_i(f, V, u) \le L_i^{\max}, T_{i+1}(f, V, u) \le T_{\max}\} \tag{12}$$

Here Π_{FV} denotes the five admissible voltage–frequency pairs in the signed policy table. The pair constraint prevents combinations outside that table from being counted as candidates; it is consistent with the twenty-candidate bound used in Section V.d. and selects the feasible point of least composite cost,

$$(f_i^*, V_i^*, u_i^*) = \underset{(f,V,u)\in\Omega_i}{\arg\min}\ J_i \tag{13}$$

If Ω_i is empty—that is, no admissible operating point satisfies both the deadline and the thermal limit—the controller enters safe mode, admits only critical tasks and reports a deadline miss for the remainder, which is the behavior observed at the highest workload intensities in Section VII.F.

C. Risk Threshold Adaptation

A static security threshold produces excessive false alarms during naturally bursty workloads, because a camera burst or an industrial alarm legitimately raises memory and DMA counters. The threshold is therefore adapted slowly from the recent distribution of normal-mode, benign-classified scores,

$$\tau_{i+1} = (1 - \eta)\tau_i + \eta(\mu_r + \xi\sigma_r) \tag{14}$$

where μ_r and σ_r are the running mean and standard deviation of the risk score over a window of epochs classified as benign, ξ sets the detection margin in standard deviations, and η = 0.02 makes the update gradual. Updating only in normal mode reduces exposure to overt anomalies, but self-classification is not ground truth. A low-score attack can still enter the update stream; slow adaptation alone does not prove resistance to poisoning. The window length, permitted threshold bounds and a trusted freezing rule are not reported and require sensitivity and evasion tests.

D. Control Algorithm

Algorithm 1 states the intended per-epoch procedure. Its cost is dominated by the evaluation of Eq. (11) over the candidate set, so the online complexity is $O(|F||U|)$ per epoch, with $|F|$ = 5 and $|U|$ = 4 in the configuration of Section VI, i.e., at most twenty cost evaluations every 5 ms. The bound assumes five paired voltage–frequency states, not an independent Cartesian product of five voltages and five frequencies. A small candidate count establishes operation count, but it does not establish execution time, synthesized area, or power overhead.

Algorithm 1: *THERMO-SAFE-RV per-Epoch Risk-Aware Thermal Control.*

Input: policy table Π = {F, V, U, Tg, T_{max}, τ_1, τ_2, β, w, λ}; deadline L^{max}
Output: selected (f*, V*, u*) and mode s
verify(firmware, accelerator_cfg, Π)
if verification fails then halt in recovery mode
τ ← τ_1; i ← 0
while node active do
  (T, P, u) ← read_sensors(); (m, c, p, d) ← read_event_counters()
  z ← β_1m + β_2c + β_3p + β_4d − β_0; r ← σ(z) [Eqs. (5), (6)]
  if r ≥ τ_2 or T ≥ T_{max} then s ← safe; restrict_DMA()
  else if r < τ and T < Tg then s ← normal
  else s ← guard; end if [Eq. (7)]
  Ω ← ∅
  for each paired (f, V) and u ∈ U admitted by s do
    L ← latency(f, u); P ← power(f, V, u) [Eqs. (1), (2)]
    T^+ ← ρT + (1 − ρ)T_{amb} + κP [Eq. (4)]
    if L ≤ L^{max} and T^+ ≤ T_{max} then add (f, V, u) to Ω
  end for [Eq. (12)]
  if Ω = ∅ then s ← safe; admit critical tasks only; log; i ← i + 1; continue
  (f*, V*, u*) ← argmin over Ω of J [Eqs. (11), (13)]
  apply(f*, V*, u*); dispatch_inference()
  if s = normal then τ ← (1 − η)τ + η(μr + ξσr) [Eq. (14)]
  log_counters(); i ← i + 1
end while

VI. EXPERIMENTAL SETUP

The evaluation uses an architecture-level simulation model of a RISC-V edge-AI node, implemented in Python, in which Eqs. (1)–(4) drive latency, power, and temperature and the controller of Algorithm 1 selects the operating point each epoch. The workload is a compact CNN inference stream with mixed sensor frames. The simulator tracks latency, energy, temperature, deadline satisfaction, and monitor quality under the reported benign and attack-stress scenarios. The purpose is to expose architectural trends before RTL implementation. No hardware traces or post-synthesis results are provided to calibrate the absolute values. Table III lists the reported parameters.

TABLE III
SIMULATION PARAMETERS.

Parameter	Reported value
Processor model	RISC-V edge core cluster with tightly coupled CNN accelerator
Frequency states F	400, 600, 800, 1000, 1200 MHz
Voltage states V	0.65–1.05 V; five levels paired with F
Accelerator utilisation U	0.25, 0.50, 0.75, 1.00
Ambient temperature T_{amb} Thermal guard / hard limit	35–45 °C 65 °C / 75 °C
Thermal coefficients ρ, κ	0.86, 4.2 °C W^{-1}
Control epoch	5 ms
Cost-weight numerical settings w_E, w_L, w_R, w_T	0.35, 0.25, 0.25, 0.15; temperature-unit convention discussed in Section V.b
Threshold adaptation η, ξ	0.02, 3.0
Attack model	Memory-burst flooding, DMA misuse, control-flow anomaly
Comparison schemes	Baseline RV, DVFS-only, Secure Monitor, thermal guard, proposed
Workloads	TinyCNN and MobileNet-style inference streams

A. Benchmark Schemes

- Baseline RV—conventional RISC-V edge inference at the highest admissible frequency, without risk-aware or thermal control.
- DVFS-only—runtime frequency scaling driven by utilization and temperature, with no intrusion score.
- Secure Monitor—runtime anomaly monitoring and response, with no thermal-aware DVFS coordination.
- Thermal guard—strict temperature control without attack-aware policy selection; its reported attack indication is based on temperature excursions.
- THERMO-SAFE-RV—the proposed integrated controller of Algorithm 1.

All schemes execute the same workload traces, thermal model and event-counter stream, so that differences arise only from the control policy. Each reported point is the mean of twenty runs with independent attack-injection seeds. Run-level values, seed identifiers, attack/benign counts, and variance are not available in the supplied record, so confidence intervals and statistical significance cannot be reconstructed from the plotted means.

B. Modeling Assumptions and Calibration Disclosure

The study doesn't rely on a cycle-accurate processor model or on measured device attributes, but rather on analytical latency, power and single-state thermal relations. Table IV separates settings reported from settings not found in the current record. The following description of offline benign-trace calibration is retained as an architectural design element, but no fitting objective, values of the coefficient, and values of the residuals are provided, so there are an incomplete independent reproduction and validation of absolute values.

Such a device specific calibration would need to have a latency and power reference trace across the admissible states, a separate benign calibration partition for counter normalization and score parameters, and held-out residuals, all of which would require significant amounts of latency and power. It should also distinguish if any of the power has already been fitted for memory, monitor and accelerator. Those reference curves are not reported here nor is there a new calibration; the numerical curves should instead be interpreted as examples of the model, not as an experimental validation of a specific RISC-V chip.

C. Security Evaluation Scope

The attack families evaluated in this paper are memory-burst flooding, DMA misuse, and control-flow anomaly, as shown in Table III. The peripheral-power abuse is part of the threat model and feature set, but there is no separate peripheral-attack result provided. Other attacks, such as mixed attacks, low-rate evasive activity, explicit monitor-disable attempts, feature ablation, and sensitivity to a threshold/window are not reported separately. The evaluation should include matched benign controls and held-out traces and per-family counts, and detection, false alarms, onset-to-response delay and deadline costs should be reported together in a strengthened evaluation.

The security rate must be evaluated on a consistent basis, e.g. per epoch labelled, attack event or attack episode. Then the attacked units are the ones used to detect, and the suggested "false" alarms are the benign units of the declared type. The underlying counts or a precise definition of the unit is not provided for the existing percentages. They are not stored as aggregate simulation summaries and are unable to define coverage, confidence bounds or similar false-positive behavior per family under all benign workloads.

TABLE IV
MODEL AND CALIBRATION INFORMATION AVAILABLE FOR INTERPRETING THE SIMULATIONS.

Component	Available information	Information not reported
Latency / power	Functional forms in Eqs. (1)–(3); frequency and voltage ranges	L_0, α_f, α_a, $Lmon$, C_{eff}, P_{stat}(V), P_{acc}, E_{mem}, E_{mon}; units and coefficient provenance
Thermal recurrence	ρ = 0.86; κ = 4.2 °C W^{-1}; epoch 5 ms; ambient 35–45 °C	Initial temperature, ambient schedule, package reference and calibration residuals
Operating states / constraints	Five paired frequency–voltage states; four utilisation levels; guard 65 °C; limit 75 °C	Exact voltage pairs, per-task deadlines, criticality rules, T_{safe} and λ values
Security score / adaptation	Four-counter logistic form; η = 0.02; ξ = 3.0	β_0–β_4, counter-event definitions, fitting method, τ_1/τ_2, update window and threshold bounds
Workload / repetitions	TinyCNN and MobileNet-style streams; twenty attack-seed runs	Exact model weights, input dataset/resolution, traces, seeds, per-run metrics and test counts
Hardware / reaction delay	Preliminary module bars and reaction-latency curves	Area/power baseline ledger, shared-block accounting, synthesis/device data and timing timestamps

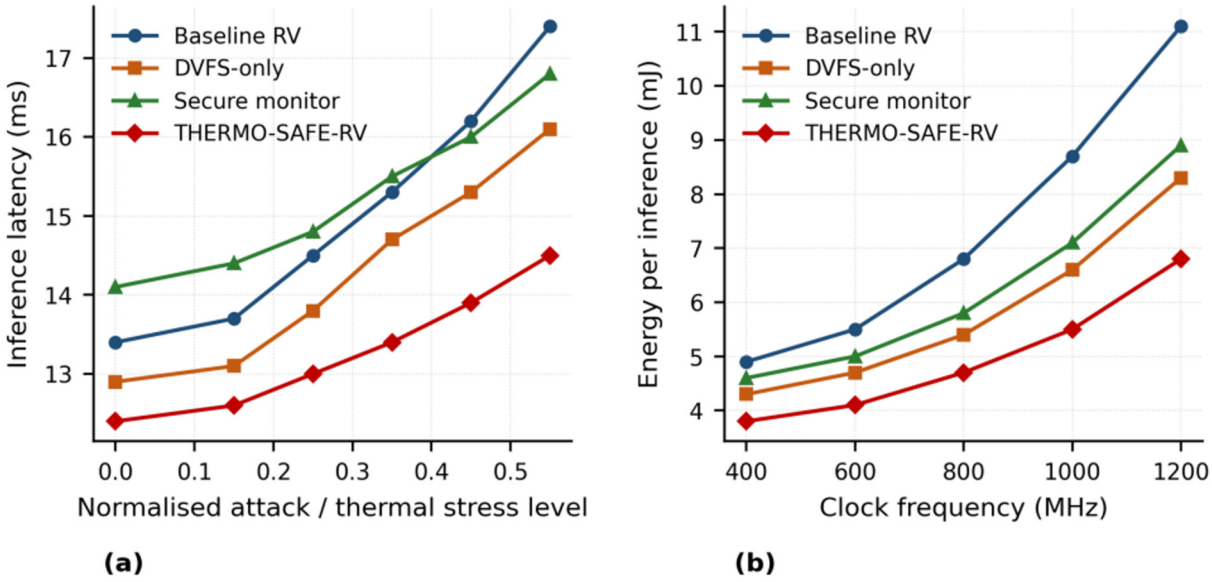


Fig. 2. (a) Inference Latency under Combined Thermal and Security Stress; (b) Energy per Inference across the Available Clock-Frequency States.

VII. RESULTS AND DISCUSSION

A. Latency Robustness and Energy Efficiency

Previous sections have shown that the adopted methodology can produce very low latency values. As the thermal and security stress becomes higher, the latency values determined by the proposed controller are the lowest among all proposed solutions, as it does not apply throttling when the RS score is low, and uses guard modes only when safety precautions are justified by either the level of risk or predicted temperature. The monitor-only solution provides the highest latency at the low level of stress, because it bears all the instrumentation costs without considering the energy-related limitations. The highest values of latency caused by the adoption of the suggested controller still provide lower latency of 14.5 ms against 17.4 ms of the baseline, thus yielding the overall decrease of 16.67% calculated based on the figures provided.

These trends are conditional on the unspecified coefficient and trace choices identified in Section VI.b. The curves do not establish execution latency or energy on fabricated hardware, and their causal interpretation cannot replace ablations or calibration against an independent reference.

B. Thermal Stability

Fig. 3 shows baseline RISC-V operation climbing steadily towards the unsafe region under sustained inference, while the proposed method stabilizes peak temperature well below the limit without sacrificing throughput. The gain comes from early guard-mode activation: because Eq. (4) predicts the temperature of a candidate operating point one epoch ahead, the controller acts before the constraint is violated. Fig. 3 is a 60.00% intensity trace; the 80.00% operating point is reported in Table 5, where the unconstrained Baseline RV exceeds the proposed controller's 75 °C limit.

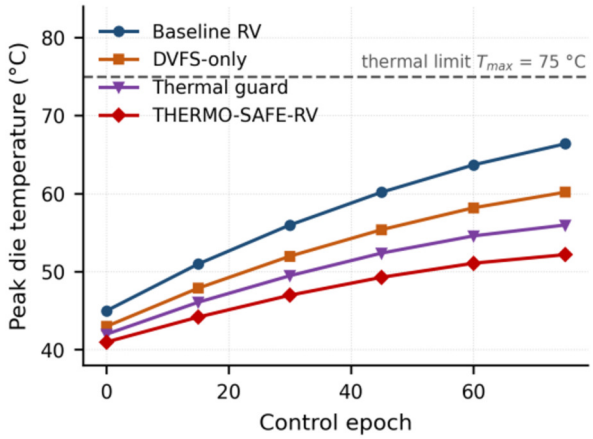


Fig. 3. Peak Die Temperature During a Sustained Workload at 60.00% Intensity, Against the 75 °C Limit.

C. Runtime Security Quality

Fig. 4 and Table V compare the reported security rates. The attack-detection increase from 91.20% to 96.80% is specifically relative to the Secure Monitor scheme, a gain of 5.60 percentage points. The corresponding false-alarm reduction from 7.80% to 3.10% is 4.70 percentage points. Baseline RV and DVFS-only have no intrusion detector; their detection and false-alarm entries are not applicable, rather than measured zero-percent performance.

The thermal-guard scheme reports 84.00% indication and 6.20% false alarms for attacks that manifest as temperature excursions. These values describe a thermal indication mechanism, whereas Secure Monitor and THERMO-SAFE-RV use runtime anomaly monitoring. Naming the mechanisms prevents an unspecified baseline comparison from mixing distinct capabilities.

The joint thermal/risk rule and gradual threshold adaptation

provide a plausible architectural explanation for the reported operating point. However, the existing aggregate rates do not independently isolate their contributions or demonstrate sufficiency of the four counters. The missing feature-level and threshold-sensitivity tests identified in Sections III.a and VI.c limit the security conclusion.

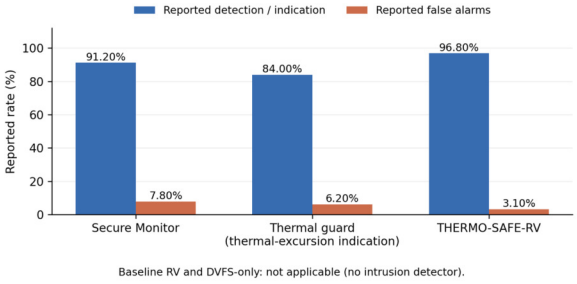


Fig. 4. Reported Attack-Detection or Thermal-Indication Rates and False Alarms. Baseline RV and DVFS-only are not applicable because no intrusion detector is implemented. Thermal guard indicates temperature excursions; Secure Monitor and THERMO-SAFE-RV use runtime anomaly monitoring.

D. Temperature Context and Classification Accuracy

The thermal guard is set at 65 °C and the maximum admissible temperature at 75 °C. At 80.00% workload intensity, Table V reports a 64.9 °C peak for THERMO-SAFE-RV, compared with 78.4 C for the unconstrained Baseline RV scheme. The 64.9 C value is a result at that operating point, not the controller limit or an independently measured classification-accuracy condition.

The earlier statement of 89.50% classification accuracy at 85 °C, compared with 81.20%, is withdrawn from the controlled-operation results because 85 °C exceeds the stated 75 °C limit and no separate out-of-envelope accuracy evaluation is documented. The associated temperature–accuracy figure is removed. No replacement accuracy value is assigned to either 64.9 °C or 75 °C. A temperature-specific accuracy claim would require the actual model, test dataset, timing/precision mechanism, and evaluation denominator; thermal control alone does not establish such accuracy.

E. Energy–Security Position and Architectural Ablation

Fig. 5(a) places each scheme in the energy–security plane. The proposed method occupies the preferred region of low energy and high composite security score, whereas the monitor-only and thermal-guard schemes each trade one axis against the other. Fig. 5(b) removes complete blocks: attestation, DVFS, the thermal guard or the full anomaly score. Removal of the anomaly score changes the reported composite score from 100 to 74, and removal of DVFS changes it from 100 to 78. These are architectural-block comparisons; they do not constitute leave-one-counter-out feature ablation or demonstrate that the selected four features are sufficient.

The normalized composite score is distinct from the attack-detection rate. Its numerical definition and uncertainty are not disclosed, so the plotted ordering is illustrative and should not be interpreted as a calibrated security probability or proof of protection against every stated attack. Independent per-feature and per-family tests remain necessary.

F. Deadline Satisfaction Under Increasing Load

Edge-AI deployments do not run at a constant input rate; camera bursts, industrial alarms, and sudden events captured by sensors lead to a higher load intensity. As can be seen from Fig. 6, the controller exhibits improved deadline satisfaction in high-load conditions owing to its evasion of reactive throttling: it uses the predicted rise in temperature, expressed through equation (4) and the context of anomaly described in equation

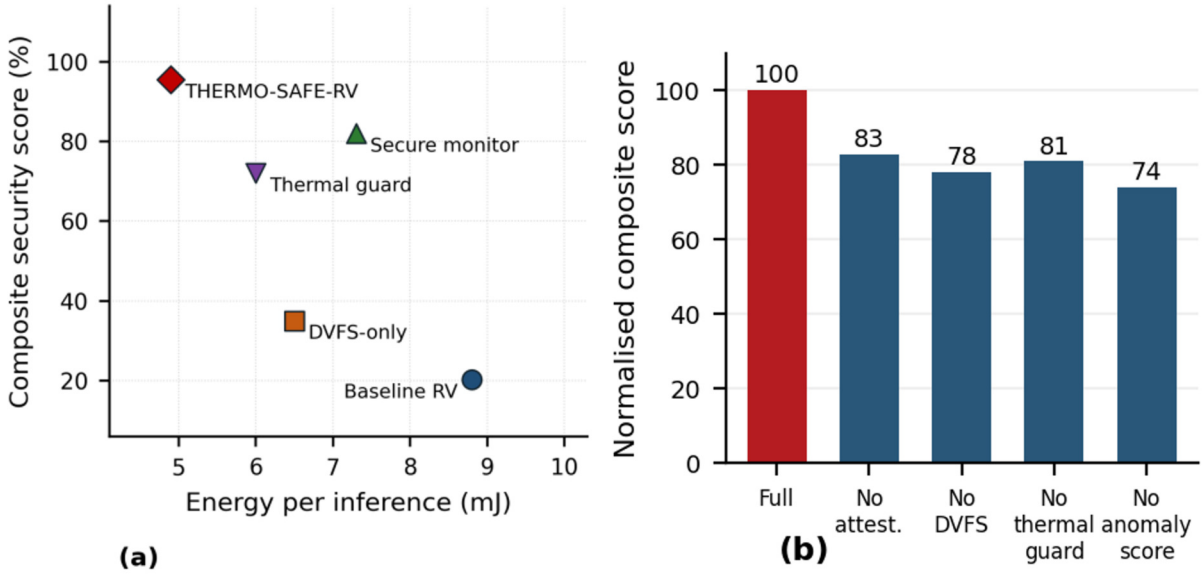


Fig. 5. (a) Energy–security Position of the Evaluated Schemes; (b) Architectural-Block Ablation, Normalised to the Full Design; Complete Anomaly Scoring Is Removed, Not Individual Counters.

(6), in order to determine the point of operation, thus ensuring that thermal limits are maintained within the next time period identified. At maximum performance, it meets deadlines in 75.00% of the cases against 47.00% for Baseline RV; the missed cases are those epochs when Ω_i is empty, allowing only critical tasks to pass.

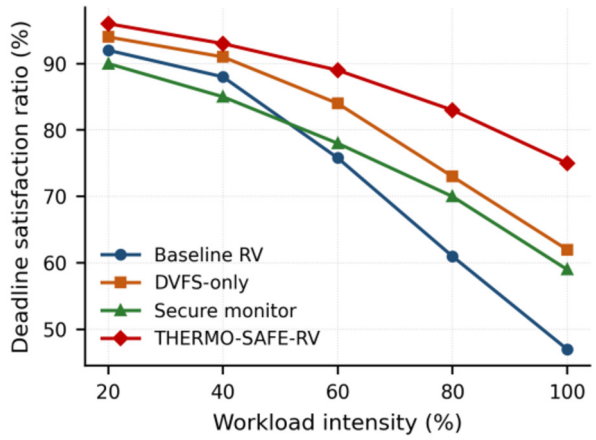


Fig. 6. Deadline Satisfaction Ratio as Workload Intensity Increases.

G. Resource Overhead and Reaction Latency

Fig. 7(a) shows preliminary area and power estimates for event counters, policy ROM, risk logic, the DVFS finite-state machine, and attestation. These are module-level estimates rather than synthesized measurements. The original aggregate cap of 0.65% area and 0.30% power is withdrawn: the plotted module contributions do not establish that total. If all five bars are non-overlapping incremental contributions relative to the same baseline, approximate readings sum to 2.20% area and 0.91% power. This conditional graphical sum is not a replacement hardware measurement; shared resources, baseline inclusion, and denominator definitions must be reconciled in a consolidated ledger.

A reproducible overhead assessment must identify baseline node area and power, added and reused hardware, technology/device, counter and policy-storage sizes, clock/voltage state, and activity assumptions. Boot-time attestation cost must be separated from runtime cost. Neither the small candidate set nor the absence of a neural detector proves a particular overhead percentage.

Fig. 7(b) reports simulation reaction latency as anomalous events per control window increase. At fifty events per window, THERMO-SAFE-RV reports 13.1 ms and Secure Monitor 20.5 ms, a 36.10% reduction calculated from those displayed values. This is a reported reaction-delay comparison, not a measured score-evaluation or controller-execution time. The 5 ms control epoch specifies sampling/decision cadence and does not establish a 5 ms reaction time.

Reaction delay should be measured from a defined attack-onset timestamp to a defined protective-action timestamp, with waiting for collection, score evaluation, policy lookup, voltage/frequency transition, and DMA enforcement distinguished. The event-window duration, onset alignment and timestamps are not supplied, so the delay components and worst-case response cannot be reconstructed. The curves therefore do not validate a hardware-visible response path.

H. Comparative Summary

Table V collects the reported architecture-level simulation values at 80.00% workload intensity. Relative to Baseline RV, THERMO-SAFE-RV reduces energy from 8.70 mJ to 5.50 mJ and peak die temperature from 78.4 °C to 64.9 °C, a decrease of 13.5 °C. The reported attack-detection

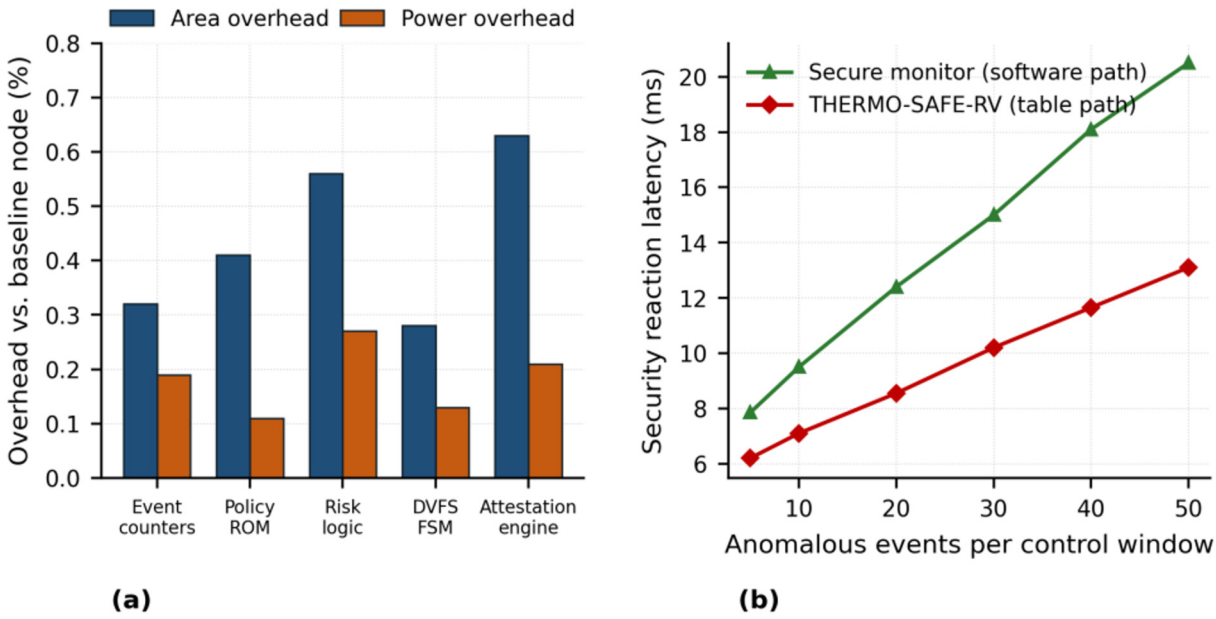


Fig. 7. (a) Preliminary Module-Level Area and Power Estimates Without a Validated Aggregate Total; (b) Reported Simulation Reaction Latency as Anomaly Density Increases. Neither panel is a synthesized or measured hardware result.

comparison uses Secure Monitor: 91.20% versus 96.80%; the false-alarm comparison is 7.80% versus 3.10%. Thermal guard is a temperature-excursion indication comparator, and Baseline RV and DVFS-only have no intrusion-detection metric.

TABLE V
SIMULATION PARAMETERS.

Method	Latency (ms)	Energy (mJ)	Peak (°C)	Detection / indication (%)	False alarms (%)
Baseline RV	16.2	8.70	78.4	N/A	N/A
DVFS-only	15.3	6.60	70.8	N/A	N/A
Secure Monitor	16.0	7.10	75.2	91.20	7.80
Thermal guard	15.1	6.10	67.4	84.00	6.20
THERMO-SAFE-RV	13.9	5.50	64.9	96.80	3.10

N/A means not applicable because no intrusion detector is implemented. Thermal-guard rates refer to temperature-excursion indication. The 78.4 °C baseline result is unconstrained by the proposed controller's policy and does not alter its 75 °C limit. The 64.9 °C value is the reported peak at this operating point, not a demonstrated maximum across all conditions.

VIII. DISCUSSION AND THREATS TO VALIDITY

A. Interpretation of the Results

Individual mechanisms are insufficient. DVFS-only control reduces energy but cannot detect hostile behavior; secure monitoring detects anomalies but raises overhead and, indirectly, temperature; thermal-guard policies protect reliability but throttle benign workloads, visible as the lowest deadline satisfaction among the thermally aware schemes. The proposed system improves the trade-off by selecting normal, guard and safe modes from a joint risk–temperature state rather than a single threshold. Fig. 5(b) is consistent with contributions from multiple architectural blocks, but its composite-score differences do not isolate the four monitored features or establish security sufficiency.

B. Limitations and Threats to Validity

First, the evaluation is architecture-level rather than post-synthesis. Latency, power, and temperature follow Eqs. (1)–(4), but the numerical workload/power coefficients, calibration references and run-level records are absent. Absolute chip performance, statistical significance and transfer to a different package or workload therefore cannot be established from the reported averages. RTL implementation and FPGA measurement remain future work.

Second, the four monitored signals cover only observable deviations at the selected interfaces. A stealthy adversary can match the benign memory, DMA, and peripheral distribution, and permitted-path data misuse may produce no control-flow alert. Peripheral-specific, mixed and low-rate attacks, matched benign controls, per-feature ablation and threshold/window sensitivity are not reported. Runtime anomaly scores do not independently prove access-control enforcement.

Third, offline calibration is described for β_1–β_4, cost settings and thresholds, but its fitting procedure, numeric coefficients and calibration/test split are not given. A workload shift requires recalibration and re-signing of the policy table, and self-classified benign adaptation can admit low-score attacks. Protected policy state, authenticated threshold bounds and enforcement of monitor configuration require implementation tests.

Fourth, module-overhead estimates lack a reconciled total and reaction-latency curves lack timestamp and transition details. The corrected mode priority in Eq. (7) must be checked at boundary conditions and in the simulator. Interpretation of the temperature coefficient in Eq. (11) must remain consistent with its units; changing to a new normalization would require recalibration and rerunning the results.

Finally, the 65 °C guard, 75 °C admissible limit and 64.9 °C simulated peak at 80.00% intensity describe different quantities. They do not establish classification accuracy at either 75 °C or 85 °C. The unsupported temperature–accuracy result is excluded, and any future accuracy study must separately document its operating envelope, model, dataset and error mechanism.

IX. CONCLUSION

This paper presented THERMO-SAFE-RV, a thermal-aware secure RISC-V edge-AI architecture for resource-constrained nodes. It combines secure-boot attestation, runtime intrusion monitoring, thermal-aware DVFS, accelerator scheduling, and policy-driven safe modes under one controller, formulated as a constrained problem with risk-dependent operating modes and a per-epoch search subject to latency and temperature limits. At 80.00% workload intensity, the reported simulation reduces energy per inference from 8.70 mJ to 5.50 mJ and peak temperature from 78.4 °C to 64.9 °C relative to Baseline RV. Compared specifically with Secure Monitor, detection increases from 91.20% to 96.80%, and false alarms decrease from 7.80% to 3.10%. The existing architectural-block ablation suggests contributions from several components; it does not validate a sufficient four-feature detector. Consolidated hardware-cost accounting, trace-based calibration, feature-level security tests and implementation measurements are required before deployment claims can be made.

X. DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[2] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[3] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proc. 44th Annu. Int. Symp. Comput. Archit. (ISCA)*, Toronto, ON, Canada, 2017, pp. 1–12.

[4] M. Horowitz, “Computing’s energy problem (and what we can do about it),” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, 2014, pp. 10–14.

[5] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, “EIE: Efficient inference engine on compressed deep neural network,” in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Seoul, South Korea, 2016, pp. 243–254.

[6] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, Apr. 2017.

[7] A. Waterman and K. Asanović, Eds., *The RISC-V Instruction Set Manual, Volume I: Unprivileged ISA*, Document Version 20191213. RISC-V Foundation, Dec. 2019.

[8] C. Celio, D. A. Patterson, and K. Asanović, “The Berkeley out-of-order machine (BOOM): An industry-competitive, synthesizable, parameterized RISC-V processor,” EECS Dept., Univ. California, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2015-167, Jun. 2015.

[9] F. Zaruba and L. Benini, “The cost of application-class processing: Energy and performance analysis of a Linux-ready 1.7-GHz 64-bit RISC-V core in 22-nm FDSOI technology,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 11, pp. 2629–2640, Nov. 2019.

[10] P. D. Schiavone, F. Conti, D. Rossi, M. Gautschi, A. Pullini, E. Flamand, and L. Benini, “Slow and steady wins the race? A comparison of ultra-low-power RISC-V cores for Internet-of-Things applications,” in *Proc. 27th Int. Symp. Power Timing Model., Optim. Simulation (PATMOS)*, Thessaloniki, Greece, 2017, pp. 1–8.

[11] M. Gautschi, P. D. Schiavone, A. Traber, I. Loi, A. Pullini, D. Rossi, E. Flamand, F. K. Gürkaynak, and L. Benini, “Near-threshold RISC-V core with DSP extensions for scalable IoT endpoint devices,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 10, pp. 2700–2713, Oct. 2017.

[12] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, “Temperature-aware microarchitecture,” in *Proc. 30th Annu. Int. Symp. Comput. Archit. (ISCA)*, San Diego, CA, USA, 2003, pp. 2–13.

[13] D. Brooks and M. Martonosi, “Dynamic thermal management for high-performance microprocessors,” in *Proc. 7th Int. Symp. High-Perform. Comput. Archit. (HPCA)*, Monterrey, Mexico, 2001, pp. 171–182.

[14] T. Chantem, X. S. Hu, and R. P. Dick, “Temperature-aware scheduling and assignment for hard real-time applications on MPSoCs,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 10, pp. 1884–1897, Oct. 2011.

[15] A. K. Coskun, T. S. Rosing, and K. Whisnant, “Temperature aware task scheduling in MPSoCs,” in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Nice, France, 2007, pp. 1659–1664.

[16] R. Sailer, X. Zhang, T. Jaeger, and L. van Doorn, “Design and implementation of a TCG-based integrity measurement architecture,” in *Proc. 13th USENIX Security Symp.*, San Diego, CA, USA, 2004, pp. 223–238.

[17] I. Anati, S. Gueron, S. Johnson, and V. Scarlata, “Innovative technology for CPU based attestation and sealing,” in *Proc. 2nd Int. Workshop Hardware Archit. Support Security Privacy (HASP)*, Tel Aviv, Israel, 2013.

[18] I. De Oliveira Nunes, K. Eldefrawy, N. Rattanavipanon, M. Steiner, and G. Tsudik, “VRASED: A verified hardware/software co-design for remote attestation,” in *Proc. 28th USENIX Security Symp.*, Santa Clara, CA, USA, 2019, pp. 1429–1446.

[19] L. Szekeres, M. Payer, T. Wei, and D. Song, “SoK: Eternal war in memory,” in *Proc. IEEE Symp. Security Privacy (S&P)*, Berkeley, CA, USA, 2013, pp. 48–62.

[20] M. Abadi, M. Budiu, Ú. Erlingsson, and J. Ligatti, “Control-flow integrity principles, implementations, and applications,” *ACM Trans. Inf. Syst. Secur.*, vol. 13, no. 1, pp. 1–40, Oct. 2009.

[21] Y. Kim, R. Daly, J. Kim, C. Fallin, J. H. Lee, D. Lee, C. Wilkerson, K. Lai, and O. Mutlu, “Flipping bits in memory without accessing them: An experimental study of DRAM disturbance errors,” in *Proc. ACM/IEEE 41st Int. Symp. Comput. Archit. (ISCA)*, Minneapolis, MN, USA, 2014, pp. 361–372.

[22] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proc. IEEE Symp. Security Privacy (S&P)*, San Jose, CA, USA, 2017, pp. 39–57.

[23] W. Wolf, “Hardware-software co-design of embedded systems,” *Proc. IEEE*, vol. 82, no. 7, pp. 967–989, Jul. 1994.

[24] G. De Micheli and R. K. Gupta, “Hardware/software co-design,” *Proc. IEEE*, vol. 85, no. 3, pp. 349–365, Mar. 1997.

[25] P. Coussy and A. Morawiec, Eds., *High-Level Synthesis: From Algorithm to Digital Circuit*. Dordrecht, The Netherlands: Springer, 2008.

[26] S. Mittal, “A survey of techniques for approximate computing,” *ACM Comput. Surv.*, vol. 48, no. 4, pp. 1–33, May 2016.

[27] V. K. Chippa, D. Mohapatra, K. Roy, S. T. Chakradhar, and A. Raghunathan, “Scalable effort hardware design,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 9, pp. 2004–2016, Sep. 2014.

[28] J. Lin, W.-M. Chen, Y. Lin, J. Cohn, C. Gan, and S. Han, “MCUNet: Tiny deep learning on IoT devices,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 11711–11722.

[29] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, “Once-for-all: Train one network and specialize it for efficient deployment,” in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, 2020.

[30] T. Chen, T. Moreau, Z. Jiang, L. Zheng, E. Yan, M. Cowan, H. Shen, L. Wang, Y. Hu, L. Ceze, C. Guestrin, and A. Krishnamurthy, “TVM: An automated end-to-end optimizing compiler for deep learning,” in *Proc. 13th USENIX Symp. Oper. Syst. Design Implement. (OSDI)*, Carlsbad, CA, USA, 2018, pp. 578–594.

[31] D. Lee, D. Kohlbrenner, S. Shinde, K. Asanović, and D. Song, "Keystone: An open framework for architecting trusted execution environments," in *Proc. 15th Eur. Conf. Comput. Syst. (EuroSys)*, Heraklion, Greece, 2020, pp. 1–16.

[32] R. Sahita, A. Patra, V. Shanbhogue, S. Ortiz, A. Bresticker, D. Reid, A. Khare, and R. Kanwal, "CoVE: Towards confidential computing on RISC-V platforms," in *Proc. 20th ACM Int. Conf. Comput. Frontiers (CF)*, Bologna, Italy, 2023, pp. 315–321.

[33] T. Tan and G. Cao, "Thermal-aware scheduling for deep learning on mobile devices with NPU," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 10706–10719, Dec. 2024.

[34] J. Huang, H. Sun, F. Yang, S. Gao, and R. Li, "Energy optimization for deadline-constrained parallel applications on multi-ECU embedded systems," *J. Syst. Archit.*, vol. 132, Art. no. 102739, Nov. 2022.

[35] H. Sobhani, S. Safari, J. Saber-Latibari, and S. Hessabi, "Realism: Reliability-aware energy management in multi-level mixed-criticality systems with service level degradation," *J. Syst. Archit.*, vol. 117, Art. no. 102090, Aug. 2021.

[36] Z. Li, S. Ren, and G. Quan, "Energy minimization for reliability-guaranteed real-time applications using DVFS and checkpointing techniques," *J. Syst. Archit.*, vol. 61, no. 2, pp. 71–81, Feb. 2015.

[37] F. Ponzina, S. Machetti, M. Rios, B. W. Denkinger, A. Levisse, G. Ansaloni, M. Peón-Quirós, and D. Atienza, "A hardware/software co-design vision for deep learning at the edge," *IEEE Micro*, vol. 42, no. 6, pp. 48–54, Nov.–Dec. 2022.

[38] P. P. Ray, "A review on TinyML: State-of-the-art and prospects," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1595–1623, Apr. 2022.

[39] J. Zidar, T. Matić, I. Aleksi, and Ž. Hocenski, "Dynamic voltage and frequency scaling as a method for reducing energy consumption in ultra-low-power embedded systems," *Electronics*, vol. 13, no. 5, Art. no. 826, Feb. 2024.

[40] B. Marshall, G. R. Newell, D. Page, M.-J. O. Saarinen, and C. Wolf, "The design of scalar AES instruction set extensions for RISC-V," *IACR Trans. Cryptogr. Hardw. Embedded Syst.*, vol. 2021, no. 1, pp. 109–136, Dec. 2020.

[41] L. Gerlach, D. Weber, R. Zhang, and M. Schwarz, "A security RISC: Microarchitectural attacks on hardware RISC-V CPUs," in *Proc. IEEE Symp. Security Privacy (SP)*, San Francisco, CA, USA, 2023, pp. 2321–2338.



Keshav Kumar is an Assistant Professor in the Department of Electronics and Communication Engineering, Pranveer Singh Institute of Technology, Kanpur, India. He has submitted his Ph.D. thesis in the field of hardware security at Lingaya's Vidyapeeth, Faridabad, India. He received the M.E. degree in Electronics and Communication Engineering, with a specialization in hardware security, from Chitkara University, Punjab, India. He has previously worked with Chandigarh University, Punjab, India, as a Junior Research Fellow (JRF) at NIT Patna, India, and as an Assistant Lecturer at Chitkara University, Punjab, India. He has authored or co-authored 7 books and more than 75 research papers, and has collaborated with researchers from 20 countries. His areas of specialization include hardware security, deep learning and machine learning techniques, green communication, low-power VLSI design, wireless sensor networks (WSN) and the Internet of Things (IoT). He has taught Python programming, embedded systems, IoT, computer networks and digital electronics. He is associated with Gyancity Research Consultancy Pvt. Ltd. and is a member of IAENG. His work has received more than 1400 citations, with an h-index of 19 (Google Scholar) and 13 (Scopus).



Mohit Kumar Srivastava received the B.Tech. degree from G. B. Pant Engineering College, Pauri Garhwal, India, the M.Tech. degree from HBTI Kanpur, India, and the Ph.D. degree in Electronics Engineering from Dr. A. P. J. Abdul Kalam Technical University, Lucknow, India. He is currently the Dean (Academics) at Pranveer Singh Institute of Technology, Kanpur, India. He has more than 25 years of academic and industrial experience, including positions at MPEC, Philips India Ltd., HCL Infosystems Ltd. and Lohia Starlinger Ltd. His current research interests include wireless communication, signal processing, filter bank multicarrier modulation and MIMO communications.



Vivek Kumar received the Ph.D. degree in Electronics and Communication Engineering from BIT Mesra, Ranchi, India, in 2018. He is currently a Professor and the Head of the Department of Electronics and Communication Engineering at Pranveer Singh Institute of Technology (PSIT), Kanpur, India, and has more than 13 years of academic, research and administrative experience. His doctoral research focused on body-worn SS-UWB antennas and IEEE 802.15.6 CSMA/CA MAC protocol modelling for medical wireless body area network applications. His research interests include wireless body area networks (WBANs), UWB and wearable antennas, wireless communication, networking protocols, visible light communication (VLC), fiber-optic sensors and healthcare IoT. He has published more than 30 research articles in Scopus- and Web of Science-indexed journals and conference proceedings, has contributed book chapters to international publishers, and holds patents in UWB antennas, healthcare systems, IoT, robotics, blockchain-based security and intelligent systems. He received the Best Teacher Award (Young Associate Professor) from Dr. A. P. J. Abdul Kalam Technical University (AKTU), Lucknow, and a Best Paper Presentation Award at ICRTC-2020.



Utkarsh Pandey is an Assistant Professor in the Department of Electronics and Communication Engineering, Pranveer Singh Institute of Technology, Kanpur, India, and is currently pursuing research in electronics. His research interests include microstrip and ultrawideband antennas, wearable and flexible antennas for wireless body area networks, communication systems for IoT, 5G and WLAN applications, and, more recently, machine learning and artificial intelligence. He has published research articles in journals including Scientific Reports, Electronics Letters and Wireless Communications and Mobile Computing, and has several published Indian patents.



Amandeep Kaur received the Ph.D. degree from Chitkara University, Punjab, India, and is with Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India. She has more than 18 years of experience in teaching, research and academic administration. She has guided several Ph.D. scholars to completion and is currently supervising doctoral research scholars. She has published numerous research papers in reputed national and international journals and conferences, and is actively engaged in research, innovation and knowledge dissemination, with a commitment to academic excellence and a research-oriented learning environment.