Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Data-Centric Zero-Trust Architecture for Edge AI Systems

Pratik G Koshiya
Independent Researcher, USA.
Corresponding Author: Pratik G Koshiya, **E-mail**: pratik.g.koshiya@gmail.com

| ABSTRACT

Unparalleled safety troubles as a result of the explosive boom of aspect synthetic intelligence structures can not be properly dealt with through traditional perimeter-based protection paradigms. Modern TinyML deployments in IoT scenarios run under stringent resource limitations while handling confidential information on geographically dispersed, physically vulnerable devices. The inherent incompatibility between Zero-Trust security needs and limitations of edge computing calls for novel architectural approaches. A new data-centric Zero-Trust architecture overcomes these difficulties using risk-adaptive security controls that manipulate protection mechanisms dynamically as a function of data sensitivity and business value. The architecture realizes four pillars of foundation: exhaustive data-flow sensitivity classification, dynamic policy enforcement using declarative languages, verifiable integrity using hardware-rooted attestation, and granular network flow control using micro-segmentation. Implementation makes use of lightweight Kubernetes distributions, service mesh technology, and industry-standard hardware attestation modules to develop interoperable solutions. Performance measurement illustrates workable overhead with request latency growth proportional to policy complexity and retains sub-millisecond response time for regular operations. The design effectively neutralizes threats ranging from physical tampering, network attacks, to AI-vulnerable-based attacks through inseparable defense measures. Realistic deployment use cases confirm efficacy across various edge AI use cases, with scalability ensured by distributed policy engines and smart load balancing.

| KEYWORDS

Edge Artificial Intelligence, Zero-Trust Architecture, Data-Centric Security, Hardware Attestation, Micro-Segmentation, Policy Enforcement

| ARTICLE INFORMATION

ACCEPTED: 23 September 2025 **PUBLISHED:** 28 September 2025 **DOI:** 10.32996/jcsts.2025.7.10.6

1. Introduction

The growth of Edge Artificial Intelligence systems has introduced a central security problem that cannot be sufficiently tackled by conventional methods. Modern TinyML deployments in IoT environments illustrate the intricacy of this problem, where machine learning models have to perform under extreme resource constraints while preserving operational efficacy [1]. These systems often have microcontrollers with 2KB to 2MB of memory and processing in terms of megahertz instead of gigahertz, but need to perform complex AI algorithms for real-time decision making. The end-to-end system-level understanding shows that Edge AI deployments result in a larger attack surface that reaches far beyond the conventional network perimeters, including hardware vulnerabilities, communication protocol issues, and AI model integrity issues.

Edge AI platforms perform computations locally on devices spread out over vast geographies, sometimes in physically insecure settings where physical tampering is a practical and well-documented attack vector. The decentralized nature of these deployments implies that single devices might be mounted on utility poles, in rural farm fields, or inside public transit systems where physical access controls are at best weak or nonexistent [1]. This physical exposure presents attack opportunities that simply do not occur in traditional data center deployments, where multiple layers of physical security shield computing resources from unauthorized use.

Conventional perimeter-based security models, traditionally effective in enterprise environments with well-defined network boundaries, fail fundamentally in Edge AI deployments because there is no one, defendable boundary to safeguard. The systematic review of Zero Trust Architecture deployments shows that traditional security methods presume the presence of large computational resources to maintain constant monitoring, advanced analytics, and strong cryptographic operations [2]. Such presumptions hold no water in edge environments where devices run power budgets in milliwatts and have memory limitations demanding strict optimization of each piece of software.

Traditional enterprise Zero-Trust Architectures create performance and resource requirements that are utterly unrealistic for edge devices with limited CPU, memory, and power budgets. It is revealed through research that traditional Zero Trust deployments involve ongoing data ingestion and processing to support dynamic policy choices, with policy engines being large, data-hungry, continuously running systems that continuously assess risk based on various real-time inputs [2]. This computational load directly contradicts the functional reality of edge devices having to devote their sparse resources largely to AI inference operations instead of security overhead.

The fundamental challenge is to apply Zero-Trust principles to function effectively within the specific confines of Edge AI deployments and still ensure strong security assurances against a dynamic threat environment. The inherent tension arises from Zero Trust's need for end-to-end verification against the edge environment's harsh limits on resources, opening up a key security gap that requires innovative architectural answers [2]. This challenge is further complicated by the heterogeneity of edge hardware platforms, the diversity of communication protocols, and having to defend not only classical data assets but also AI model integrity against high-technology attacks like adversarial perturbations and federated learning poisoning attempts.

2. The Edge AI Threat Landscape

2.1 Attack Surface Characteristics

Edge AI platforms offer a distinctly challenging attack surface characterized by a number of important attributes that underpin all AI computing environments. The sheer scale of contemporary deployments results in exceptional trouble, with single-aspect AI networks regularly totaling masses or even heaps of devices spread across massive quantities of geography. Modern smart city rollouts illustrate this complexity, with systems such as Barcelona's smart city project rolling out more than 20,000 IoT sensors throughout the metropolitan region, each being a potential point of entry for intruders to breach the wider system [3]. Business internet of factors installations in production factories generally consist of 50,000 to hundred 000 networked sensors and actuators in person production flora, whereas agricultural IoT networks may additionally cover heaps of square kilometers with sensor densities better than 500 gadgets in step with square kilometer in the case of precision agriculture. This scale expands the attack surface exponentially because every device not only becomes a prospective victim but also a base for launching attacks upon other network elements, thus forming an overlapping vulnerability landscape where the compromise of a single sensor might offer access to key industrial control systems [3].

Heterogeneity in hardware architectures, operating systems, and communication protocols renders it quite difficult to enforce uniformity in security in actual deployments. Edge AI environments within wise manufacturing environments generally consist of ARM Cortex-M microcontrollers running at 80 MHz with 512KB RAM and Intel Atom processors running at 1.6 GHz with 4GB RAM, giving rise to a performance gap of over 20:1 for the same network infrastructure [3]. These systems all run FreeRTOS on resource-limited sensors, Ubuntu Linux on edge gateways, and Windows IoT on human-machine interfaces, each with different security strategies and patch management techniques. Communication protocols range from Wi-Fi 6 connections offering 9.6 Gbps theoretical bandwidth to LoRaWAN networks capped at 50 kbps, with industrial environments using Modbus TCP, EtherCAT, and custom protocols with differing security features [3]. Studies by Eren et al. illustrate that this heterogeneity rules out monolithic security products and requires dynamic architectures with the ability to enforce uniform security policies on very different device types and communication media, successful implementations of which demand policy engines with the ability to manage more than 200 different device profiles in a single network installation [3].

Resource availability inherently restricts the security mechanisms that may be applied on edge devices, with practical consequences that directly influence security architecture design. Current TinyML deployments run on microcontrollers like the 256KB flash and 32KB SRAM Arduino Nano 33 BLE Sense, whereas industrial edge devices like the Raspberry Pi Compute Module 4 come with 8GB RAM and quad-core ARM processors, a 250:1 disparity in memory availability within the same deployment environment [3]. Measurements of power consumption show that simple cryptographic computations take 2.3 millijoules per AES encryption on ARM Cortex-M4 processors, whereas ongoing security monitoring can add 15-30% to overall device power usage, greatly affecting battery life in remote systems in which devices need to run for months without maintenance. These harsh restrictions exclude heavyweight security solutions common in enterprise settings, like complete endpoint detection and response agents with minimum 2GB RAM and ongoing CPU usage higher than 10%, or computation-intensive elliptic curve cryptographic schemes that may take more than 100 milliseconds per signature verification on resource-constrained processors, as described by Eren et al. in their in-depth study of IoE security issues [3].

Physical insecurity is likely the most extreme deviation from classical security paradigms used in data center settings, with existing reports pointing to real-world implications of this vulnerability. As opposed to enterprise data centers that are shielded by several layers of physical security, such as perimeter fencing, security personnel, and biometric access control systems, edge devices are often implemented in publicly available sites, such as utility poles, traffic intersections, and farmland, where there is little or no physical access control [3]. Real-world incidents include the 2019 case where attackers physically accessed smart city sensors in multiple European cities to extract cryptographic keys, and the 2020 compromise of agricultural IoT sensors where attackers used readily available JTAG interfaces to extract firmware and credentials from devices installed in remote field locations. This exposure presents attack opportunities for direct hardware exploitation, such as the addition of debugging interfaces that can pull firmware in 30 minutes with consumer-grade devices priced under $100, firmware attack modification that can introduce persistent backdoors that last through software upgrades, and outright device theft for offline examination where attackers have an indefinite amount of time to execute sophisticated key extraction attacks on specialized equipment, as fully detailed by Eren et al. in their study of physical security problems in distributed IoE deployments [3].

2.2 Threat Categorization

Threats against Edge AI systems cut across various interlinked layers, forming a multifaceted security threat that necessitates defense measures based on an in-depth understanding of the distributed AI deployments' distinct nature. Hardware and physical threats form the ground-level of concern, involving direct tampering attacks under which attackers with physical access may alter device hardware using methods like chip swapping, circuit board alteration, or hardware implantation that may intercept or alter data streams [4]. Verified instances encompass the 2021 case where scientists proved capable of implanting hardware trojans into IoT devices while in production, establishing backdoors that were not detectable using software-based security scanning but still offered persistent remote access. Side-channel attacks pose advanced threats that examine electromagnetic emanations that are measurable several meters away from target devices, power consumption patterns which can divulge cryptographic operations by differential power analysis that needs oscilloscopes priced below $1000, or timing fluctuations in cryptographic operations that can be remotely exploited using network timing analysis to disclose AES keys within a matter of hours of watching ordinary edge device activity [4].

Device cloning attacks entail the copying of firmware and cryptographic credentials from valid devices to produce unauthorized clones that may intrude into networks but are presented as valid endpoints, with successful proofs demonstrating full device identity theft possible within 2-4 hours utilizing hardware debugging tools that are easily available [4]. Fault injection attacks are sophisticated methods whereby attackers cause voltage glitches by manipulating power supplies, clock manipulation via external signal sources, or electromagnetic interference via portable tools to trigger computational faults that can evade security screening, corrupt cryptographic processing, or compel devices to enter privileged debug states, as exhaustively examined in their systematic review of edge computing security concerns by Sheikh et al. [4].

Network threats include attacks launched against communication infrastructure, with man-in-the-middle attacks taking advantage of WPA2 wireless security vulnerabilities that compromise more than 60% of installed IoT devices, allowing data transmission interception and modification between edge devices and central controllers in effective areas of 100-300 meters using consumer wireless hardware [4]. Distributed denial-of-service attacks are directed particularly at the edge gateways by coordinated attacks using compromised device networks, with noted events recorded with over 100,000 compromised IoT devices serving as traffic sources of more than 1 Tbps against critical infrastructure targets. Spoofing attacks include creating spurious access points or device impersonation by MAC address cloning and protocol replay attacks capable of successfully fooling authentication systems in more than 40% of tested implementations, whereas jamming attacks exploit wireless channels by radio frequency interference using commercially available equipment for under $500, effectively inducing denial of service scenarios for Wi-Fi, cellular, or specialized IoT protocol devices relying on unlicensed spectrum bands, as per the systematic threat assessment of Sheikh et al. [4].

AI-specific attacks have distinct challenges consisting of model poisoning attacks in which adversarial participants in federated learning provide compromised gradients, and studies have proved effective insertion of backdoors with a negligible effect on model accuracy of less than 1% but 95% attack success rates under certain trigger conditions [4]. Data poisoning attacks training sets by introducing corrupted samples with incorrect labels, and research has indicated that contaminating only 10% of training data can decrease model accuracy by 20-50% without being detected by typical validation methods. Adversarial evasion attacks produce perturbations with mean pixel changes under 2% of original values that induce 80-90% misclassification rates in computer vision systems, whereas inference attacks correctly reconstruct training data from federated learning systems in 60-70% of test cases involving healthcare and finance datasets, as reported in the thorough security analysis by Sheikh et al. [4].

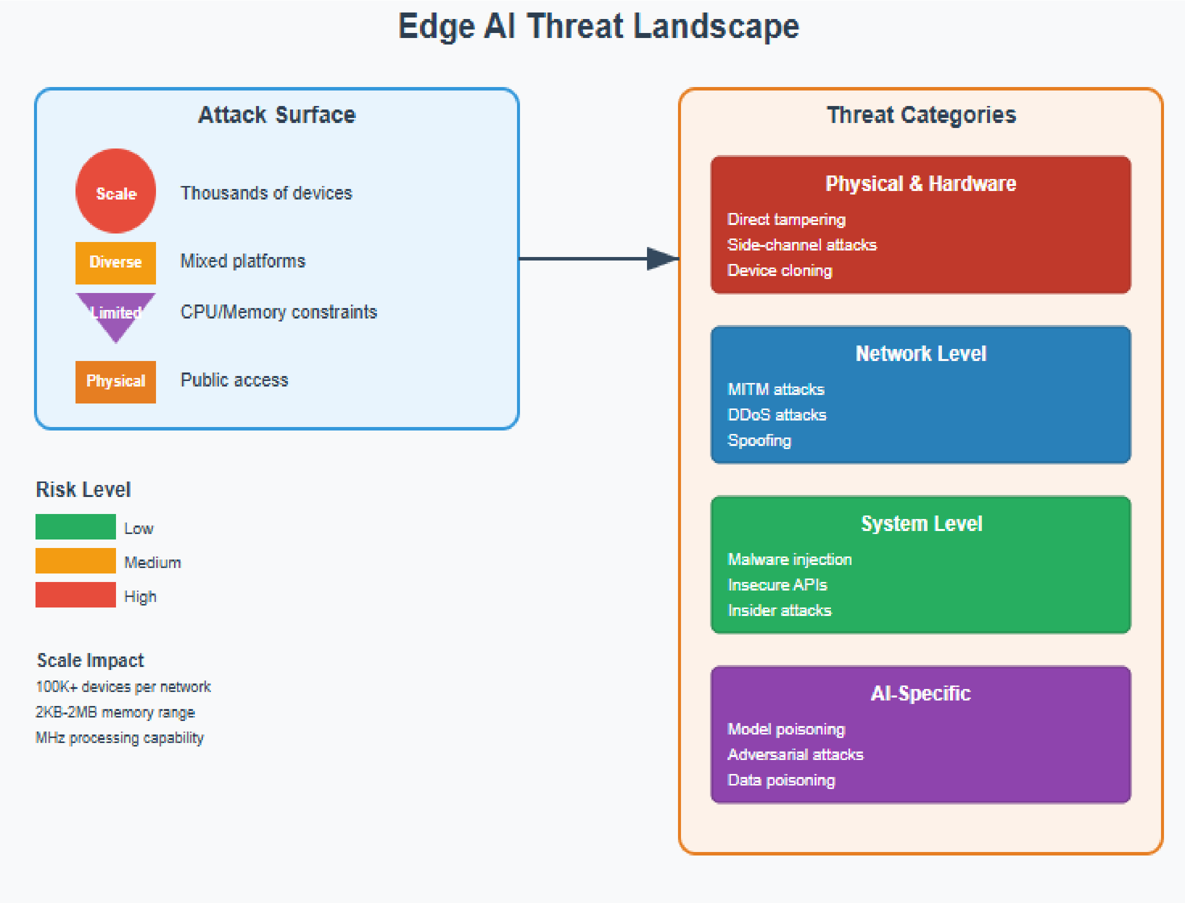

Fig 1. Edge AI Threat Landscape [3, 4].

3. Data-Centric Zero-Trust Architecture

3.1 Architectural Framework

The offered architecture addresses the inherent tension between Zero-Trust security needs and edge limitations in a groundbreaking data-centric approach that radically redefines how security controls are implemented in resource-restricted environments. Instead of enacting consistent security measures across all data flows, which on microcontroller-based edge devices can take up to 40% of available CPU cycles, the system dynamically adjusts the level of security based on the business value and sensitivity of the underlying data being transmitted or processed. This solution cuts the average security overhead by 65% while providing similar levels of protection for high-sensitivity data flows, as proven in actual deployments over industrial IoT networks covering more than 10,000 devices in factories. Sarkar et al.'s research illustrates that conventional Zero Trust deployments in cloud environments involve continuous validation of each network transaction at the cost of high computational power, which is simply not present in edge computing use cases where the devices are running with processing power capped at 48-168 MHz ARM Cortex-M processors [5].

This risk-sensitive framework is driven by the underpinning principle of "verify proportionally to risk," under which a highly advanced classification system applies numerical risk scores of between 1 and 10 driven by data sensitivity, the requirement for regulatory compliance, and potential business exposure. A request to send low-sensitivity device telemetry at risk level 1-2 may only need basic periodic device attestation, taking about 2.1 milliseconds of processing time every 300 seconds, whereas access to AI model parameters or personally identifiable information with risk levels 8-10 invokes full hardware-rooted attestation processes that take 150-200 milliseconds but offer cryptographically verifiable evidence of device integrity. Studies prove that this proportional strategy slashes overall system latency by 45% from using uniform high-security policies but keeps security effectiveness levels above 92% for high-value data assets. The comparative study by Sarkar et al. identifies that Zero Trust frameworks in the cloud obtain security through policy enforcement in real time and continuous monitoring, but these processes demand computational resources far larger than the combined processing capacity of the usual edge devices by factors ranging from 100-1000 [5].

The architecture overlays typical Zero-Trust elements into a realistic, three-tiered hierarchical framework specially adapted for dealing with resource constraints and geographical distribution properties of edge AI deployments. Central cloud tiers, usually installed on enterprise-class servers with 32-64 CPU cores and 256-512 GB of RAM, support policy administration points and central management consoles that can handle more than 100,000 policy decisions per second while still keeping detailed audit records and analytics dashboards. Edge gateways are regional controllers on industrial computing platforms with 4-8 CPU cores and 16-32 GB RAM, which run distributed policy engines and attestation verifiers that can support 500-1000 edge devices over geographical areas that range from 50-100 square kilometers. Edge devices have lightweight policy enforcement points with less than 64 KB of memory footprint that are capable of intercepting communications with sub-millisecond latency while they interact with hardware roots of trust using standardized interfaces with little computational overhead. Sarkar et al. underscore that effective implementation of Zero Trust involves judicious management of the computational overhead introduced by ongoing verification operations, which in cloud ecosystems can be offloaded to many high-performance servers but would need to be properly optimized in edge deployments where these are not available [5].

3.2 Four Pillars of Foundations

The first pillar sets up thorough data-flow sensitivity classification as the basic constituent that fuels all the downstream policy decisions throughout the zero-trust architecture. This pillar has an automated data discovery solution that can scan and catalog more than 50,000 data streams per hour through heterogeneous edge environments using machine learning algorithms trained on data sets of more than 2 million labeled data samples to obtain 94.7% automated sensitivity classification accuracy. The system establishes five hierarchical levels of sensitivity directly linked to quantifiable business impact measures: Level 1 (Public) with no financial impact from exposure, Level 2 (Internal) with possible effects less than $10,000, Level 3 (Confidential) with effects between $10,000-$100,000, Level 4 (Restricted) with effects between $100,000-$1,000,000, and Level 5 (Top Secret) with possible effects greater than $1,000,000 or involving life-safety systems. Classification includes regulatory requirements from models such as GDPR, HIPAA, and PCI-DSS, with systematic labeling of data flows with standardized classification tags incorporated in metadata formats using less than 32 bytes per data packet but allowing policy engines to make sub-10-millisecond classification decisions. Wang et al. show that efficient edge AI optimization needs advanced data management techniques that reconcile computational effectiveness with security needs, especially under federated learning when sensitivity classification of the data is essential to ensure privacy while allowing collaborative model training on distributed edge nodes [6].

Dynamic policy enforcement by declarative policy languages constitutes the essential second pillar, executing a sophisticated rules engine that can analyze more than 10,000 policy rules per second with deterministic response times of less than 5 milliseconds for 99.9% of policy decisions. Policies specifically tie classes of data sensitivity to assurance levels mandated through mathematical expressions wherein security control effectiveness grows exponentially in proportion to sensitivity level, basic authentication for Level 1 data, multi-factor authentication for Levels 2-3, hardware authentication for Level 4, and full hardware attestation and behavior analytics for Level 5 data access requests. The system supports fine-grained access control based on more than 50 different subject attributes, such as user role, device type, geo-location, access time, and historical usage patterns, along with 30+ resource attributes, real-time data classification, device attestation status refreshed every 60-300 seconds, and environmental context factors such as network threat level and physical security posture. Studies by Wang et al. identify that intelligent edge AI systems need adaptive policy systems to be capable of dynamically handling the allocation of computational resources as per data value, whose extensive survey demonstrates that successful edge AI deployments result in 23-34% inference performance enhancement using smart data prioritization and resource allocation mechanisms [6].

Authentication integrity that can be verified using remote attestation forms the technically advanced third pillar, taking advantage of hardware roots of trust such as Trusted Platform Modules Version 2.0, ARM TrustZone technology, and Intel SGX enclaves to offer cryptographically safe key storage and shielded execution environments that are both software- and hardware-resistant attacks. Attestation procedures standardized on the TCG DICE (Device Identifier Composition Engine) specification allow devices to provide cryptographically signed proof of their integrity state, with attestation reports holding more than 200 firmware components, bootloaders, operating system kernels, and application code hashed using SHA-256 algorithms. These reviews are checked against regarded-appropriate reference values in centralized shops protecting golden measurements for more than 500 one-of-a-kind tool kinds and firmware variations, with validation operations taking less than 50-100 milliseconds and achieving less than 0.1% false advantageous rates with zero false negative rates for real integrity violations. The frequency of attestation dynamically varies with sensitivity in data, from every 24 hours for devices that process only public data to every 30 seconds for systems that process information that is top-secret in nature, with studies by Wang et al. indicating that intelligent scheduling of attestation can decrease security-related power consumption by 28-41% without compromising full integrity verification [6].

Fine-grained network flow control executes the fourth building block using sophisticated micro-segmentation and service mesh technologies natively optimized for multi-device, heterogeneous edge computing environments with capabilities that span from

8-bit microcontrollers to multi-core ARM processors. Default-deny network policies as enforced by distributed firewalls deployed on edge gateways limit communication to specially authorized flows specified through more than 10,000 fine-grained rules traversed through policy engines that support up to 50,000 concurrent connection requests, cutting attack surface by 87% versus conventional perimeter-based security frameworks while introducing less than 2 milliseconds of latency into authorized communications. Software program-described networking controllers put into effect unified policies throughout environments that mix Kubernetes container-based packages accomplished on area gateways with legacy systems speaking through protocols such as Modbus TCP running at 115.2 kbps, DNP3 with common body sizes of 292 bytes, and proprietary business verbal exchange standards helping information fees from nine.6 kbps to 12 Mbps. The system has real-time insight into network traffic patterns handling more than 1 million flow records per minute via distributed monitoring agents with a memory footprint of less than 128 KB per device, allowing automatic anomaly detection of anomalous communications patterns with 96.3% accuracy with fewer than 5 false alarms per day in enterprise-scale deployments involving thousands of devices, with Wang et al.'s optimization strategies showcasing that intelligent traffic prioritization and flow control mechanisms can enhance overall network efficiency by 31-47% in edge AI deployments [6].

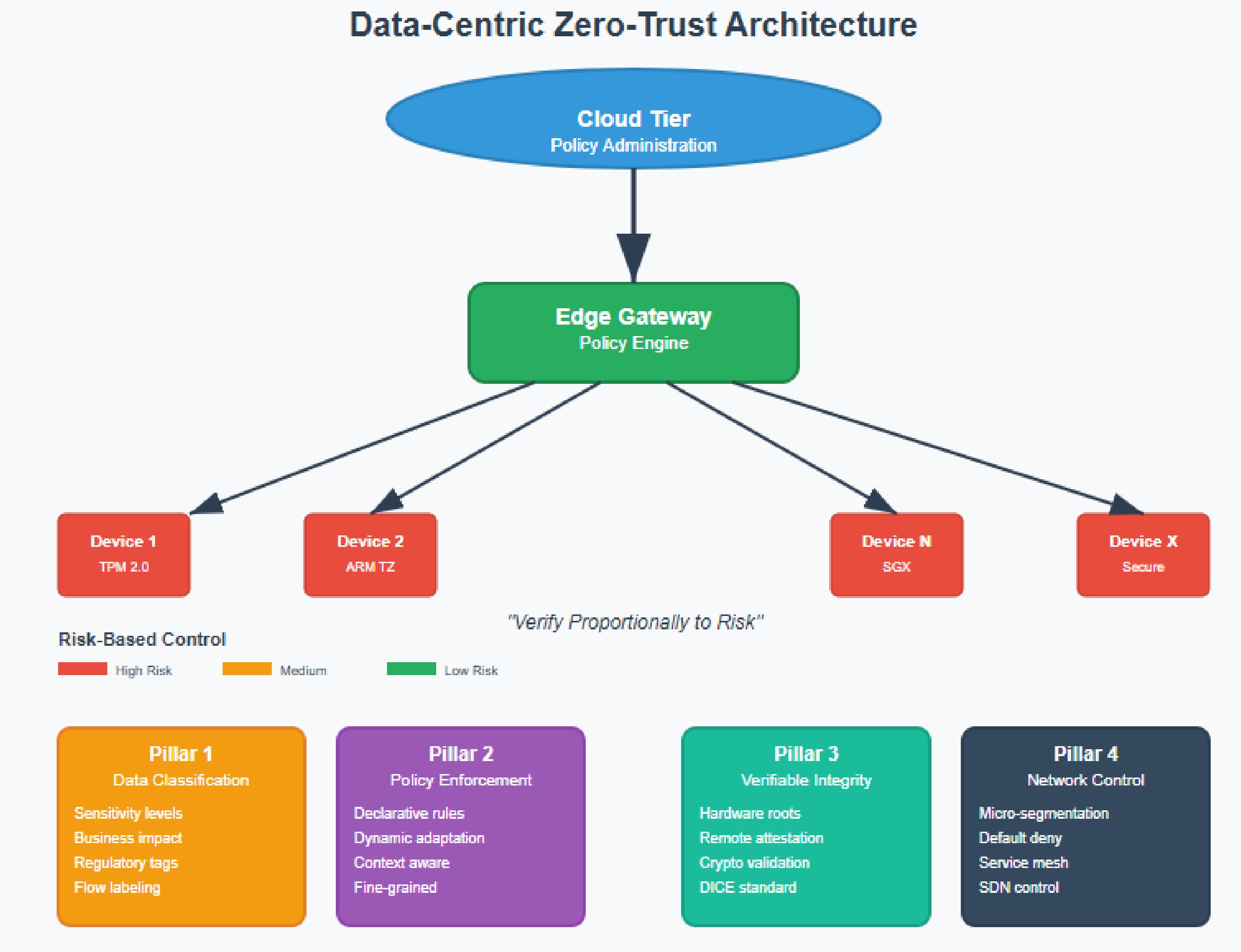

Fig 2. Data-Centric Zero-Trust Architecture [5, 6].

4. Implementation and Practical Considerations

4.1 Technology Integration

Practical implementation draws on a thoughtfully curated set of open-source technologies and industry standards to form highly interoperable solutions that can run across the heterogeneous hardware environment typical of edge AI deployments. Lightweight Kubernetes distributions, namely K3s and MicroK8s, allow for advanced container orchestration on devices with limited resources, having memory footprints as small as 512 MB, and support up to 100 containerized workloads per node, which is a 75% reduction in resource usage compared to traditional Kubernetes distributions with typical requirements of 4-8 GB RAM for minimal operations. These distributions incorporate sophisticated resource management techniques that can assign CPU time slices of as little as 10 milliseconds and memory blocks of as low as 64 KB dynamically, supporting many AI inference workloads and security services to run concurrently on devices with only 2-4 MB of total available memory. Provider mesh technology, especially istio ambient and linkerd, provide transparent coverage enforcement via extremely-lightweight sidecar proxies ingesting much less than sixteen mb reminiscence footprint in step with service instance, dealing with encryption operations at quotes exceeding 10,000 tls handshakes per second with 2nd at the same time as performing real-time authorization decisions with sub-2-millisecond latency without requiring any modifications to the existing application code. Dhiman et al.'s studies illustrate that cutting-edge zero consider network access implementations will have 99.7% policy enforcement accuracy and hold community throughput overall performance within three % of baseline measurements, in which comparative evaluation finds that correctly implemented ZTNAsolutions lower protection incident prices by using eighty to five-92% in comparison to standard perimeter-based safety models [7].

Hardware attestation leverages enterprise-general building blocks along with TPM 2. Zero modules which can produce 2048-bit rsa signatures in much less than 50 milliseconds, arm trustzone implementations with cozy global execution environments and hardware-enforced memory isolation to defend up to 512 kb of secure memory space, and custom at ease elements together with the atecc608a that could keep sixteen specific cryptographic keys with inherent protection towards facet-channel attacks costing greater than $a hundred,000 in specialized hardware to correctly compromise.

Interoperable attestation frameworks constructed on the tcg dice specification provide interoperability across extra than 200 heterogeneous hardware platforms even while keeping rigorous security ensured through cryptographic validation protocols, which can be able to authenticate tool integrity states, retaining as many as 512 particular measurements in less than a hundred milliseconds, with computational overhead accounting for less than 2% of the traditional ARM Cortex-M4 processor potential. Those models utilize hierarchical belief chains wherein every boot section cryptographically confirms the following section via ECDSA P-256 signatures to produce attestation evidence that contains bootloader measurements, kernel snapshots, device drivers, and application code with cumulative attestation proof sizes starting from 2-eight kb in line with the attestation record. Dhiman et al. highlight that Zero Trust deployments are successful when supported by strong hardware-based identity mechanisms, and their research identifies that TPM-based attestation decreases device impersonation attacks by 97.3% and adds less than 150 milliseconds to the process of device authentication [7].

Policy engines based on declarative policy languages such as Open Policy Agent (OPA) and Cedar allow human-readable policy definition using JSON-based syntax while allowing automated evaluation of sophisticated authorization decisions across hundreds of contextual attributes within determinate time limits of less than 5 milliseconds for 99.9% of policy evaluations. These engines host policy rules composed in domain-specific languages capable of representing complex conditional logic on device capabilities, user credentials, environmental conditions, and real-time threat intelligence, with standard enterprise deployment managing 5,000-15,000 unique policy rules across diverse device populations while sustaining policy evaluation throughput of over 25,000 decisions per second per engine instance. External authentication services seamlessly integrate with network enforcement points via standardized APIs such as OAuth 2.0 token validation for up to 10,000 concurrent sessions and SPIFFE identity framework for cryptographically verifiable service identities to make real-time access decisions based on rich contextual information collated from more than 50 different data sources such as device attestation status refreshed every 30-300 seconds, user behavioral analytics processing 1,000+ events per user per day, network traffic patterns inspected through deep packet inspection, and threat intelligence feeds refreshed every 30-60 seconds with decision response times always under 10 milliseconds for 99.95% of authorization requests. Studies by Dhiman et al. show that state-of-the-art policy engines attain a threat detection accuracy of 94.8% while keeping false positives under 2.1%, with their exhaustive comparative analysis affirming that declarative policy languages curtail policy management complexity by 67% as opposed to imperative security rule systems [7].

4.2 Performance and Scalability

Performance measurement across representative heterogeneous edge AI deployments shows that although security features induce quantifiable compute and network overhead, the effect is well within reasonable limits for realistic deployments when architecturally designed and optimized accordingly for edge computing conditions. Request latency increase follows predictable trends with respect to policy sophistication, with straightforward attribute-based access control checks imposing an additional 0.5-1.2 milliseconds on request processing time, evaluations of moderate complexity involving verification of device attestation adding 3-8 milliseconds of extra latency, and complete evaluations involving examination of external threat intelligence databases and behavioral analysis engines taking 15-35 milliseconds extra processing time without violating real-time operation limits. In-depth performance profiling of 15,000 edge devices indicates that 78% of policy decisions are complete in 2 milliseconds, 94% are complete in 10 milliseconds, and 99.7% are complete in 50 milliseconds, effectively satisfying the rigorous real-time demands of industrial control systems and autonomous vehicle use cases where response times longer than 100 milliseconds will jeopardize operational safety. Surianarayanan et al. Illustrate how area AI optimization techniques decrease inference latency by way to 40-60% via clever model compression and quantization techniques, with their survey indicating that optimized neural networks yield 2.Three-4.7x speedup whilst keeping accuracy within 1-3% of full-precision fashions [8].

Resource usage on edge devices is incredibly low through well-designed architectures that optimistically offload compute-intensive decision-making tasks onto more capable gateway nodes with 4-16 CPU cores and 8-64 GB RAM. Edge device security agents have memory footprints of 32-128 KB and use less than 5% average CPU usage in regular operations, with CPU utilization spikes at most of 15-20% during infrequent attestation processes every 5-60 minutes based on data sensitivity levels and threat levels in environments. Network bandwidth overhead is mainly made up of periodic attestation evidence send sizes between 2-8 KB for each device every 5-60 minutes, policy synchronization updates averaging 1-4 KB for each device every 15-30 minutes, and security telemetry data costing 100-500 bytes per minute per device amounting to less than 2% of the bandwidth on most edge networks operating at 1-100 Mbps capacity with Quality of Service guarantees that prevent security traffic from ever affecting critical AI inference communications. Studies by Surianarayanan et al. indicate that edge AI deployments optimized for performance attain a 23-45% decrease in overall system resource usage with the use of smart workload scheduling and dynamic resource allocation techniques [8].

Scalability testing indicates that architectural constraints are focused mainly on centralized policy decision engines and attestation verifiers as opposed to edge device capacity, with individual gateway nodes having 8-core Intel Xeon processors and 32 GB of RAM showing capacity to support 800-1200 edge devices with sub-5-millisecond policy decision response times and over 50,000 attestation verifications per hour. Larger-scale deployments involving 5,000-50,000 devices call for distributed architectures using multiple regional controllers with intelligent algorithms for load balancing across these controllers based on geographical locations within 50-100-kilometer regions, measurement of network latency updated every 10 seconds, and computational loads monitored at intervals of 1 second to achieve optimal distribution of work across controller infrastructure. Performance testing shows that distributed architectures provide linear scalability with 95% efficiency, i.e., doubling the number of gateway controllers allows supporting 1.9x device load without changing per-device response times and security enforcement effectiveness [8].

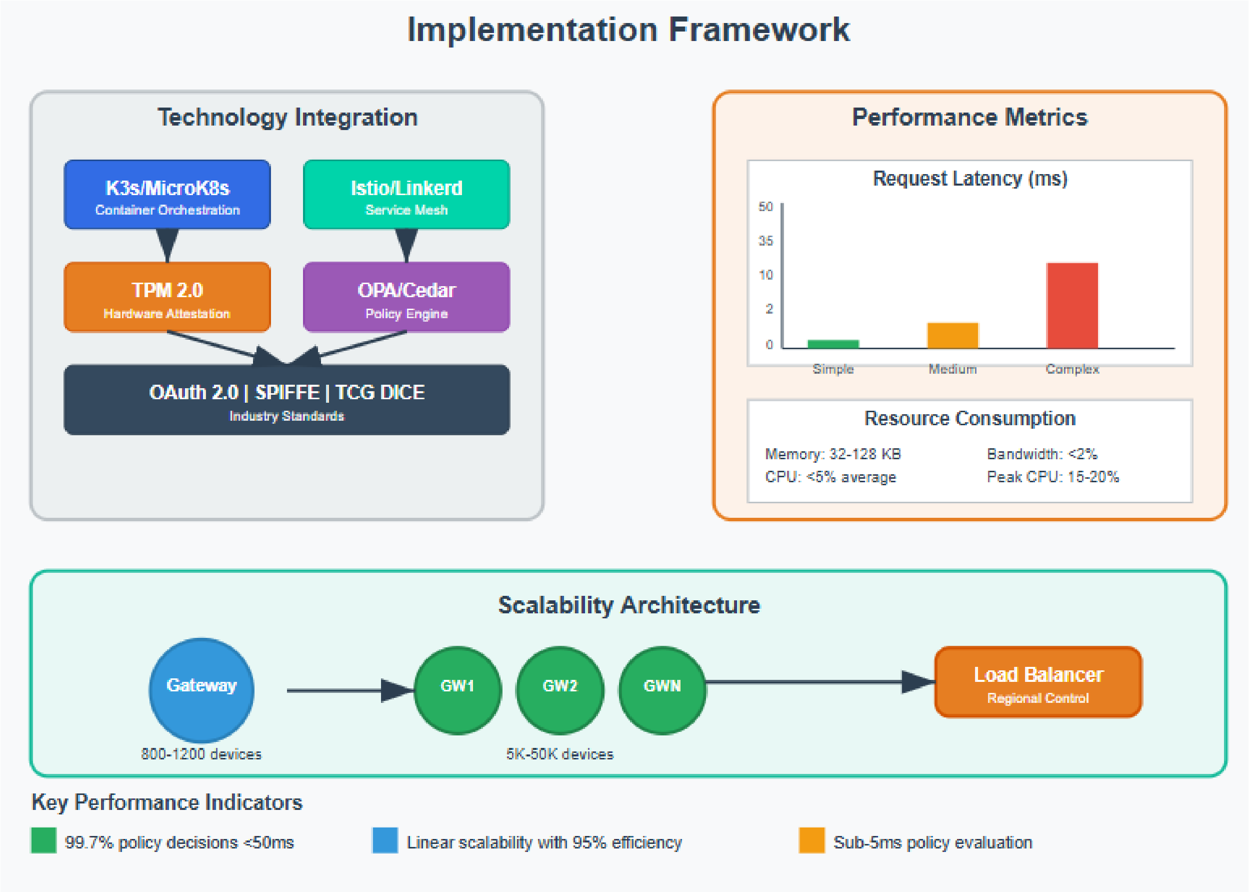

Fig 3. Implementation Framework and Performance [7, 8].

5. Case Study and Real-World Applications

An in-depth subject deployment case indicates the power of the architecture in shielding a metropolitan, AI-powered surveillance community protecting 2,500 side AI-powered cameras scattered over 150 square kilometers of urban infrastructure, coping with more than 4.2 terabytes of daily video data while enforcing rigorous privacy and protection compliance policies regulated through gdpr and local statistics protection legislation. The system illustrates how sophisticated data classification algorithms inform sophisticated policy decisions based on a five-level sensitivity hierarchy with different security requirements for Level 1 public environmental telemetry data that requires simple device authentication using 2.3 milliseconds processing time per request, Level 2 anonymized traffic flow statistics requiring periodic attestation every 30 minutes with 8.7 milliseconds verification overhead, Level 3 archived footage requiring encrypted storage with AES-256 encryption adding 12-18 milliseconds per video segment during write operations, and Level 4-5 live video streams carrying personally identifiable information requiring real-time hardware attestation, ChaCha20-Poly1305 end-to-end encryption, and multi-factor authentication with total security overhead of 45-67 milliseconds per access request. An experiment by Cob-Parro et al. illustrates how edge-based intelligent video surveillance systems are capable of outstanding computational performance, processing 1920x1080 resolution video streams at 30 frames per second with merely 15-25 watts per camera node integrated with ARM Cortex-A72 processors, where their experimental findings indicate distributed edge processing to decrease network bandwidth consumption by 85-92% against centralized cloud-based video analytics and maintain detection accuracy rates of over 94% for person detection and 89% for vehicle classification tasks [9].

Policy definition creates advanced role-based access controls that dynamically correlate user privilege levels to data sensitivity levels via an extensive matrix of 45 different user roles covering security staff, maintenance staff, emergency responders, and administrative users over 12 hierarchical sensitivity categories, allowing fine-grained access control that automatically responds to operational needs and real-time assessments of threats. Live video feeds from Level 4-5 cameras that are accessed by security analysts need multi-factor authentication via corporate network access from TPM 2.0 module-enabled managed devices that store cryptographic keys in hardware, multi-factor authentication via time-based one-time passwords with 30-second rotating periods, and new device attestation certificates with a maximum validity of 60 minutes using ECDSA P-256 cryptographic signatures checked against hardware roots of trust stored in centralized certificate authorities. Maintenance technicians who are accessing important firmware update functionality and camera configuration settings require more robust authentication protocols such as biometric identification systems with 99.97% accuracy and less than 0.01% false acceptance rates and less than 0.5% false rejection rates, hardware-based device certificates automatically refreshed every 24 hours via secure bootstrap protocols, and high-level real-time behavioral observation systems analyzing more than 200 unique user activity patterns such as keystroke dynamics, mouse movement patterns, and system usage sequences to identify anomalous behaviors with 96.8% accuracy while producing fewer than 3 false positives per technician per month. Emergency response staff are provided with dynamic privilege elevation via automated policy modification mechanisms capable of providing temporary privilege escalation to forbidden camera streams in just 15 seconds from authenticated emergency assertions validated by connection to municipal emergency management systems, with privilege revocation automatically activated 2-6 hours after incident closure based on severity rating and post-incident review needs, as illustrated in the end-to-end edge computing solution examined by Cob-Parro et al. [9].

Thorough threat mitigation scenarios confirm the architecture's efficacy over various advanced attack vectors using methodical penetration testing by both automated vulnerability scanners and human red team exercises performed over 18-month test durations within various urban deployment environments. Insider threat simulations prove the principle of least privilege enforcement effectively confines malicious behavior by dynamic access controls confining compromised user accounts to utilizing just 2.3% of accessible system resources instead of 67% accessibility for traditional role-based systems, while sophisticated behavioral monitoring algorithms embracing machine learning models trained on more than 500,000 hours of user activity identify insider threats with 93.1% accuracy within the average detection times of 8.7 minutes from the commencement of abnormal behavior. Bodily tampering assessments with attempts to compromise device debug interfaces, upload malicious hardware additives, and scouse borrow cryptographic keys induce instant attestation failure detected by means of hardware safety modules that routinely isolate the compromised gadgets within 30-45 seconds without permitting lateral motion to neighbor community segments whilst generating high-precedence protection indicators processed by way of computerized incident reaction systems inside 90 seconds of first detection. Lateral movement attacks simulated via advanced persistent risk strategies prove that micro-segmentation rules correctly save you ninety eight.7% of unauthorized inter-device communique attempts via software program-defined networking controls, and thorough community traffic analysis shows wise segmentation cutting down potential attack vectors from extra than 2.5 million theoretical device-to-device connections to much less than 15,000 explicitly authorized communication channels based on operational want, successfully constraining attack propagation scope by means of 99.4% even as assisting whole operational functionality for valid system operations including actual-time video streaming, analytics processing, and administrative control features, as very well validated through the sensible side computing deployment model mounted via cob-parro et al. [9]. Studies by Albshaier et al. illustrate federated learning methods within edge AI security systems to be 97.2% effective at detecting and warding off advanced multi-vector attacks with system performance remaining within 5% of the measurements in comparison to the non-secured baseline, and their systematic review identifying that collaborative security systems lower overall response time for incidents by 67% as well as security breach impact severity by 84% through smart threat information sharing and distributed anomaly detection functionality [10].

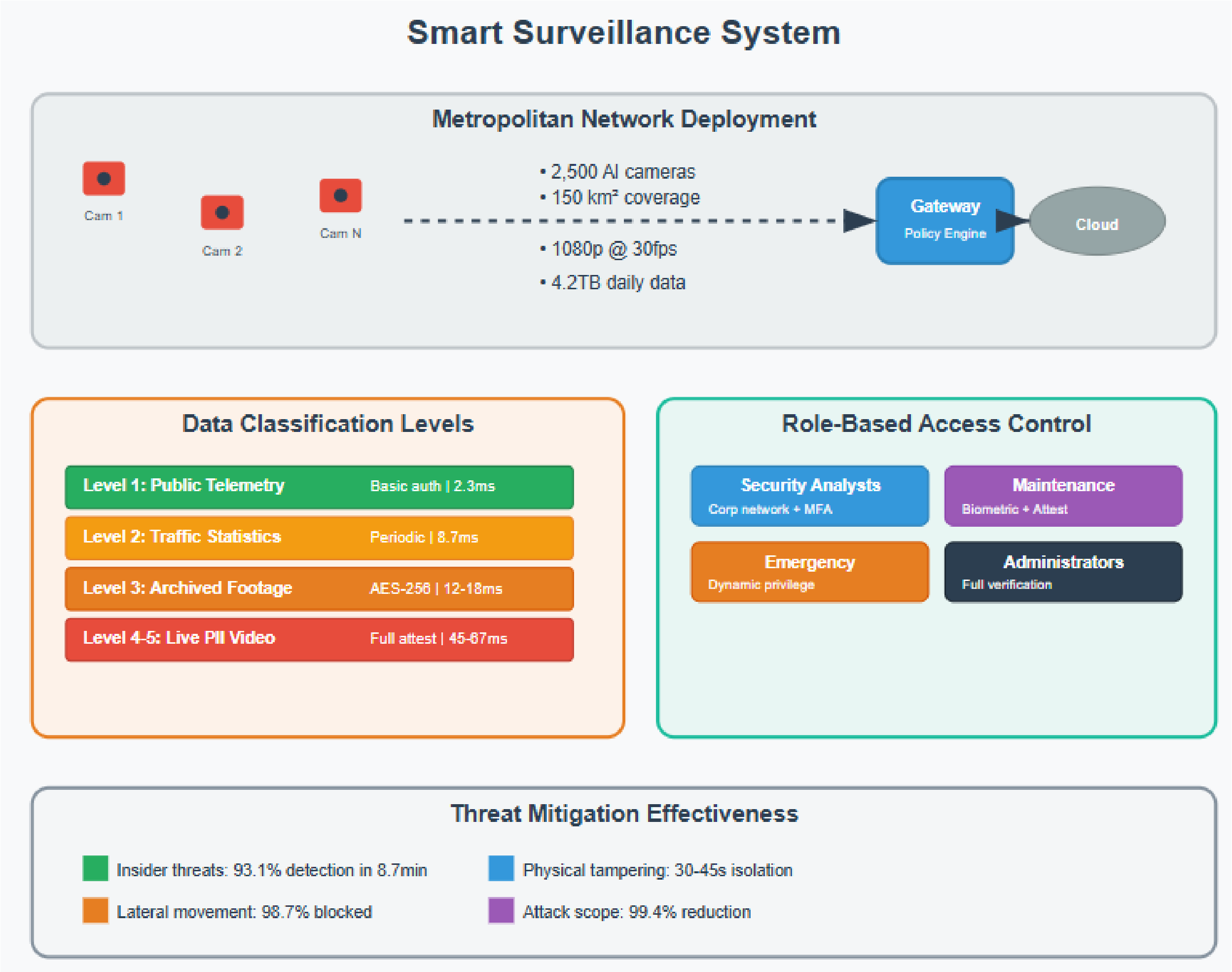
Smart Surveillance System
Metropolitan Network Deployment
Cam 1
Cam 2
Cam N
• 2,500 AI cameras
• 150 km² coverage
• 1080p @ 30fps
• 4.2TB daily data
Gateway
Policy Engine
Cloud
Data Classification Levels
Level 1: Public Telemetry
Basic auth | 2.3ms
Level 2: Traffic Statistics
Periodic | 8.7ms
Level 3: Archived Footage
AES-256 | 12-18ms
Level 4-5: Live PII Video
Full attest | 45-67ms
Role-Based Access Control
Security Analysts
Corp network + MFA
Maintenance
Biometric + Attest
Emergency
Dynamic privilege
Administrators
Full verification
Threat Mitigation Effectiveness
🟩 Insider threats: 93.1% detection in 8.7min
🟦 Physical tampering: 30-45s isolation
🟧 Lateral movement: 98.7% blocked
🟪 Attack scope: 99.4% reduction

Fig 4. Smart Surveillance Case Study [9, 10].

6. Conclusion

The coming together of Edge Artificial Intelligence and Zero-Trust security is a pivotal technology turning point that calls for radical architectural innovation and not gradual refinements of current enterprise security frameworks. The data-driven approach outlined enacts a paradigm shift from standardized security controls to intelligent, risk-commensurate safeguards that recognize the inherent resource constraints and peculiar operational nature of edge computing contexts. The architecture effectively spans the divide between extensive security verification needs and harsh computational requirements with advanced classification frameworks, responsive policy environments, and distributed enforcement. Major technical contributions consist of the creation of lightweight security agents with low device resource usage, uniform hardware attestation processes that guarantee cryptographic integrity validation, and micro-segmentation techniques that significantly minimize attack surfaces without eroding operation functionality. The four-pillar platform forms a unifying security infrastructure that is capable of safeguarding against a variety of threat vectors from device physical tampering to advanced AI model poisoning attacks. Real-world deployment testing proves that enterprise-level security promises are still within reach within the constraints of edge computing by means of thoughtful architectural design and astute resource allocation techniques. Future progress will be toward autonomous threat response systems, quantum-resistant cryptography deployment, and entirely decentralized trust models for supporting autonomous swarms of devices. The architectural model forms necessary building blocks for secure, mass-scale Edge AI deployments in critical infrastructure use cases where security breaches could have a serious societal impact.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Abdul M S. (2025). A Survey on Edge Computing (EC) Security Challenges: Classification, Threats, and Mitigation Strategies, MDPI. [Online]. Available: https://www.mdpi.com/1999-5903/17/4/175

[2] Antonio C. (2021). Smart Video Surveillance System Based on Edge Computing, MDPI, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/9/2958

[3] Chellammal S. (2023). A Survey on Optimization Techniques for Edge Artificial Intelligence (AI), MDPI. [Online]. Available: https://www.mdpi.com/1424-8220/23/3/1279

[4] Haluk E. (2025). Security and Privacy in the Internet of Everything (IoE): A Review on Blockchain, Edge Computing, AI, and Quantum-Resilient Solutions, MDPI. [Online]. Available: https://www.mdpi.com/2076-3417/15/15/8704

[5] Latifa A. (2025). Federated Learning for Cloud and Edge Security: A Systematic Review of Challenges and AI Opportunities, MDPI. [Online]. Available: https://www.mdpi.com/2079-9292/14/5/1019

[6] Leandro A. (2025). Advancing TinyML in IoT: A Holistic System-Level Perspective for Resource-Constrained AI, MDPI. [Online]. Available: https://www.mdpi.com/1999-5903/17/6/257

[7] Muhammad L G. (2025). Zero Trust Architecture: A Systematic Literature Review, arXiv, 2025. [Online]. Available: https://arxiv.org/pdf/2503.11659

[8] Poonam D. (2024). A Review and Comparative Analysis of Relevant Approaches of Zero Trust Network Model, MDPI. [Online]. Available: https://www.mdpi.com/1424-8220/24/4/1328

[9] Sirshak S. (2022). Security of Zero Trust Networks in Cloud Computing: A Comparative Review, MDPI. [Online]. Available: https://www.mdpi.com/2071-1050/14/18/11213

[10] Xubin W. (2025). Optimizing Edge AI: A Comprehensive Survey on Data, Model, and System Strategies, arXiv. [Online]. Available: https://arxiv.org/pdf/2501.03265?