Per-Device Behavioral Anomaly Detection in IoMT Networks Using Autoencoders and EVT-Based Adaptive Thresholds

Evelin E. D. Limeira[1], Rafael Roque[1], Luciano Cabral[1,3], Fernando Aires[1,2], Wellison R. M. Santos[1,2], Milton Lima[1]

[1]Centro Integrado de Segurança em Sistemas Avançados (CISSA), CESAR School, Recife, Brazil

[2]Departamento de Computação, Universidade Federal Rural de Pernambuco (UFRPE) Recife, Brazil

[3]Campus Jaboatão dos Guararapes, Instituto Federal de Pernambuco (IFPE) Jaboatão dos Guararapes, Brazil

{eedl,rrs3,lsc,faal,wrms,mvml}@cesar.org.br,

fernandoaires@ufrpe.br, luciano.cabral@jaboatao.ifpe.edu.br

Abstract. *Intrusion Detection Systems (IDSs) for the Internet of Medical Things (IoMT) must handle highly heterogeneous device traffic, which directly impacts anomaly detection and false-positive rates. This study evaluates per-device benign-only autoencoders combined with Extreme Value Theory (EVT)-based adaptive thresholds, localized reconstruction errors scoring, and SHAP-based feature attribution. Experiments are conducted on offline traffic traces from camera-class devices in the CICIoMT2024 dataset, considering behavioral, MQTT-based, and volumetric attack groups. Results indicate that per-device modeling is more effective at characterizing behavioral deviations than volumetric flooding attacks in this scenario. Furthermore, localized reconstruction-error scoring provides a more sensitive operating point than conventional global thresholding strategies.*

1. Introduction

The Internet of Medical Things (IoMT) extends networked computing into healthcare environments by interconnecting medical and monitoring devices with gateways, brokers, and cloud services. These systems support continuous observation and remote operation, but they also produce device-dependent traffic profiles shaped by heterogeneous hardware, protocols, and usage patterns. For network-level intrusion detection, this heterogeneity is not merely contextual: it directly affects benign-behavior modelling, anomaly scoring and false-positive generation. In this setting, security monitoring must preserve healthcare-data confidentiality and service availability.

Signature-based IDSs detect known attack patterns, whereas anomaly-based IDSs target previously unseen threats. Anomaly-based approaches can identify deviations from established benign behavior, particularly when labeled attack examples are unavailable [Khraisat et al. 2019, Yacoubi et al. 2026]. Anomaly detection in IoMT nevertheless remains sensitive to device heterogeneity, threshold calibration, and alert interpretability. A single model trained across different devices may hide device-specific traffic regularities.

Similarly, reconstruction-error thresholds directly affect false-positive behavior but are often treated as secondary implementation choices. Finally, reconstruction-based alerts require attribution mechanisms to support analyst triage and operational interpretation.

Despite extensive research on anomaly-based IDSs for IoT and IoMT environments, comparatively fewer studies isolate the interaction between device-specific benign modeling and threshold-calibration strategies under controlled experimental conditions.

This study investigates whether per-device benign-only autoencoders improve network-flow anomaly detection for camera-class IoMT devices when compared with a heterogeneous global autoencoder. The study uses CICIoMT2024 and evaluates four devices under the same preprocessing pipeline, architecture, training protocol, threshold-calibration procedure, and scoring policies. In this paper, *per-device* denotes one autoencoder for each named camera-device partition in the dataset, not one model for the camera class as a whole, per protocol, or per behavioral cluster. The camera subset provides four distinct profiles, enough benign traffic for training, validation, and test partitions, and the same 18 evaluated attacks for every device. The evaluation explicitly separates two design dimensions: per-device versus global training and scalar EVT-calibrated RMSE scoring versus localized reconstruction-error (LRE) scoring. SHAP-based attribution is used to identify which reconstructed features contribute most to anomaly scores.

The contribution of this work is an empirical controlled evaluation of the conditions under which per-device benign modeling is advantageous in this dataset and where its benefits are limited. The main contributions are:

- An evaluation of per-device benign-only autoencoders across four camera-class IoMT devices from CICIoMT2024, compared against a heterogeneous global autoencoder baseline;
- A four-way ablation study separating training composition from anomaly-scoring strategy, with detection rate, false-positive rate, F1, and threshold-independent AUROC evidence reported by attack subgroup;
- A paired statistical analysis across device–attack–repetition evaluation units using five repeated random holdout partitions, Wilcoxon tests, bootstrap confidence intervals, and effect sizes;
- A SHAP-based attribution analysis for behavioral protocol attacks, highlighting recurrent yet device-dependent feature-attribution patterns;
- An execution-cost characterization covering single-flow latency, batch throughput, on-disk model size, and the aggregate storage overhead of maintaining four per-device models.

The remainder of the paper is organized as follows: Section 2 presents background and related work. Section 3 describes the methodology. Section 4 reports and discusses the results. Section 5 summarizes threats to validity. Section 6 concludes the paper and outlines future work.

2. Background and Related Work

2.1. Benign-Only Detection and Threshold Calibration

Reconstruction-based anomaly detectors can be trained only on samples that represent normal operation. In this setting, autoencoders learn to reconstruct the benign training distribution, and reconstruction discrepancies are used as anomaly evidence at test

time [Sakurada and Yairi 2014]. This design avoids the need for attack samples during training, but it produces anomaly decisions instead of explicit attack labels and depends on how well the benign data represent normal operation.

Device-specific modelling is motivated by evidence that network-traffic characteristics can distinguish IoT devices and their communication profiles [Sivanathan et al. 2019]. CICIoMT2024 likewise contains heterogeneous devices and protocol environments [Dadkhah et al. 2024]. These observations support evaluating device-specific normality models, but they do not establish that specialization will outperform pooled training under every composition or attack category.

A continuous reconstruction score must be converted into an operational alert through threshold calibration. Empirical percentiles, Gaussian assumptions, and robust statistics provide comparatively simple calibration rules, whereas Extreme Value Theory (EVT) models the upper tail of the score distribution [Siffer et al. 2017]. This study uses Generalized Pareto Distribution fitting, but does not compare EVT against percentile, Gaussian, or median-absolute-deviation thresholds and makes no claim of superiority over those alternatives.

A scalar reconstruction score can hide deviations concentrated in individual variables. Localized reconstruction approaches retain feature or channel-level errors to support detection and interpretation of such deviations [Gorman et al. 2023]. The Localized Reconstruction Error (LRE) procedure used here calibrates one threshold per feature and flags a flow when any feature-level error exceeds its threshold. This logical union can increase the aggregate flow-level false-positive rate relative to each nominal per-feature rate.

SHAP provides additive feature attributions for a model output [Lundberg and Lee 2017]. In this study, SHAP is applied to feature-level reconstruction errors to identify variables associated with LRE anomaly scores. The resulting attributions describe associations with the model output and are not interpreted as causal evidence about the attack or device behaviour.

2.2. IoMT Intrusion Detection and Study Positioning

Table 1 positions this study in relation to works selected for their methodological proximity, without implying direct performance comparability. It includes IoMT studies and adjacent IoT, 5G, smart-grid, and network-security research when these address relevant gaps in threshold calibration, explainability, resource-aware deployment, or statistical validation. Differences in datasets, threat models, and evaluation protocols preclude direct ranking.

Studies using CICIoMT2024 have examined conventional classifiers, deep architectures, feature-reduced ensembles, SHAP-supported boosting, and ransomware-oriented fingerprinting [Dadkhah et al. 2024, Kavkas and Yildiz 2025, Ramesh et al. 2024, Sohail et al. 2024, Hernandez-Jaimes et al. 2024]. These approaches are predominantly supervised and therefore depend on attack classes represented during training. Benign-only methods address a different operational setting by learning normal behavior without labeled attacks, although they generally provide anomaly decisions instead of attack-type labels [Damasceno et al. 2025].

Adjacent-domain studies complement the IoMT literature by examining contrastive autoencoders, explainability, constrained deployment, efficient inference, and alternative

Table 1. Positioning of this study against related IDS works for IoMT and adjacent IoT scenarios

Reference	Technique	Dataset	Per-Device	Threshold Focus	XAI	Statistical Validation
Dadkhah et al.	Multiple ML	CICIoMT2024	No	Limited	No	No
Kavkas and Yildiz	DNN + LSTM	CICIoMT2024	No	Limited	No	No
Ramesh et al.	RF + XGBoost	CICIoMT2024	No	Limited	No	No
Sohail et al.	Boosting ensemble	CICIoMT2024	No	Limited	SHAP	No
Hernandez-Jaimes et al.	Nilsimsa + ML	CICIoMT2024	No	Limited	No	Friedman
Cavalcante et al.	DT on-device	CICIoT2023	No	Limited	No	No
Vieira et al.	Contrastive AE	CICIDS2017	No	Gaussian	No	No
Damasceno et al.	MLP + AE	5G-NIDD	No	Limited	No	No
Emelianova et al.	KAN	CICIoT2023	No	Limited	No	No
This work	**AE + EVT**	**CICIoMT2024**	**Yes**	**EVT/GPD**	**SHAP**	**Paired tests + repeated holdout**

detection architectures [Vieira et al. 2025, Quincozes et al. 2024, Cavalcante et al. 2024, Simioni et al. 2024, Emelianova et al. 2025]. Their inclusion reflects methodological relevance, not equivalence between domains. The present combines per-device benign-only training, EVT calibration, two anomaly-scoring operating points, SHAP attribution, and paired analysis at the device-attack-repetition level. It does not claim general superiority, but evaluates whether device-specific reconstruction changes detection behavior under controlled scoring conditions.

3. Methodology

This section describes the experimental protocol used to evaluate per-device benign-only anomaly detection for camera-class IoMT traffic. The study follows a controlled offline design: all models use the same preprocessing pipeline, autoencoder architecture, training protocol, threshold-calibration procedure, and test partitions. The two main design factors are training composition (per-device versus global) and anomaly-scoring mode (EVT-based global reconstruction error versus localized reconstruction-error scoring). Attack flows are used only at test time and are never used for training or threshold calibration. The central hypothesis is that device-specific benign modeling yields more stable reconstruction behavior for behavioral attacks than heterogeneous global modeling, particularly under localized anomaly scoring.

3.1. Study Design and Research Questions

The evaluation addresses five questions:

- **RQ1:** When and by how much does per-device training improve detection over a global model under the same scoring mode?
- **RQ2:** What is the detection–false-positive trade-off between EVT and LRE?
- **RQ3:** Which reconstructed features contribute most to anomaly decisions?
- **RQ4:** Are the observed effects stable across five repeated random holdout partitions?
- **RQ5:** What model-execution and on-disk storage costs are observed for the persisted per-device models?

3.2. Dataset, Scope, and Attack Taxonomy

The experiments use CICIoMT2024 [Dadkhah et al. 2024]. Four camera-class devices are evaluated: Owltron Camera, Ecobee Camera, M1T Camera, and Blink Camera. In this study, *per-device* denotes one separately trained model for each named device partition. Cameras were selected to provide several profiles within the same functional class, sufficient benign traffic for separate training and calibration, and test traffic for all 18 evaluated attacks. Other device classes are outside the study scope.

The attacks are organized into four operational subgroups, shown in Table 2, and the same taxonomy is used for subgroup-level reporting and paired statistical analysis.

Table 2. Attack taxonomy used in the evaluation (18 attack types, 4 subgroups).

Subgroup	#	Attack types
Behavioral Protocol Attacks	6	ARP Spoofing; MQTT-Malformed Data; Recon-OS Scan; Recon-Ping Sweep; Recon-Port Scan; Recon-VulScan
MQTT Connect Floods	2	MQTT-DDoS-Connect Flood; MQTT-DoS-Connect Flood
MQTT Publish Floods	2	MQTT-DDoS-Publish Flood; MQTT-DoS-Publish Flood
Volumetric Network Floods	8	TCP/IP DDoS variants (ICMP, SYN, TCP, UDP); TCP/IP DoS variants (ICMP, SYN, TCP, UDP)

Benign flows for each camera are randomly divided at the flow level into training, validation, and test partitions using a 70/15/15 ratio. The partition sizes are 16,414/3,517/3,518 for Owltron, 10,304/2,208/2,208 for Ecobee, 6,556/1,404/1,406 for M1T, and 5,523/1,183/1,184 for Blink. The heterogeneous global partitions contain 109,876/23,545/23,546 benign flows. The validation partition is used after training to compute descriptive validation MSE and calibrate EVT and LRE thresholds. Attack flows are retained exclusively for testing.

In addition to the main partition, five random train/validation/test partitions generated with different seeds are evaluated. These partitions are repeated holdouts and are not mutually exclusive folds of a conventional k-fold procedure. The manuscript reports partition-dependent attack results but does not use them to infer a specific source of variation beyond split sensitivity. The experiment does not enforce temporal, capture-level, or session-level isolation; accordingly, its results characterize flow-level generalization.

3.3. Architecture Overview and Device Routing

Figure 1 summarizes the training, calibration, detection, and attribution workflow.

In the offline experiment, each flow is routed to the corresponding model using the ground-truth device label supplied by the dataset partition. The study does not implement online identification from MAC addresses, IP addresses, switch ports, or other identifiers. The reported per-device results accordingly assume correct attribution. Identity spoofing, reassignment, and flow misrouting are not evaluated.

3.4. Preprocessing and Model Training

The original 45 flow features undergo two filtering stages. Five features with variance below 10^{-6} are removed, followed by correlation filtering at $|\rho| > 0.95$, which removes 12 additional features and leaves 28 active features. A `RobustScaler` is fitted only on the

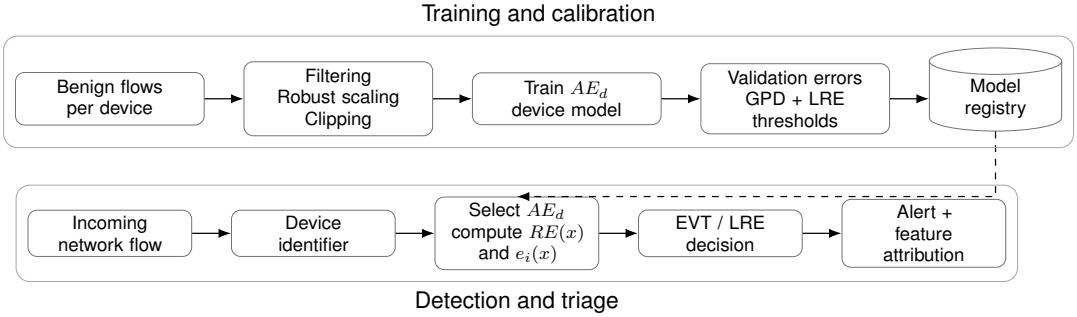

Figure 1. Per-device anomaly-detection workflow, separating training/calibration from detection/triage.

benign training partition and applied to validation and test data; scaled values are clipped to $[-5, 5]$.

For each IoMT device d, a dedicated regularized autoencoder AE_d is trained exclusively on benign traffic from that device. The architecture is symmetric:

$$\text{Input}(28) \to 256 \to 128 \to 256 \to \text{Output}(28).$$

The autoencoder is configured with the specified hidden dimensions, epoch count, batch size of 256, random seed, contamination parameter, and execution device. Training minimizes MSE using the Adam optimizer with a learning rate of 10^{-3}, together with batch normalization and dropout at a rate of 0.2. The model is fitted exclusively on the benign training partition. Validation data are not used for training, early stopping, or model selection; they are processed only after fitting to compute validation MSE and calibrate the anomaly thresholds.

As the hidden layers are wider than the input, the architecture is overcomplete and does not enforce dimensional compression. No compressive bottleneck or architecture ablation is evaluated, so the effect of this design on identity mapping and attack reconstruction is treated as a limitation.

Table 3. Training and threshold-calibration summary. Validation MSE is a post-training reconstruction metric; EVT decisions and AUROC use sample-level RMSE. FPR intervals are 95% Wilson binomial intervals computed from benign-test FP and TN counts.

Device	Training Flows	Validation MSE	EVT Threshold	FPR EVT [95% CI]	FPR LRE [95% CI]
Owltron Camera	16,414	0.00407	0.1824	1.17% [0.86, 1.58]	6.28% [5.53, 7.13]
Ecobee Camera	10,304	0.00479	0.2221	0.95% [0.62, 1.45]	6.57% [5.61, 7.68]
M1T Camera	6,556	0.00446	0.1833	1.42% [0.92, 2.19]	7.75% [6.47, 9.27]
Blink Camera	5,523	0.00370	0.1457	1.10% [0.64, 1.87]	4.73% [3.66, 6.09]
Global Baseline	109,876	0.00440	0.1709	0.94% [0.82, 1.07]	8.95% [8.59, 9.32]

The heterogeneous global reference uses the same architecture and procedure but is trained on a larger benign pool containing profiles beyond the four evaluated camera partitions. This global pool is larger than the union of the four reported per-device partitions; it is therefore interpreted as a broader heterogeneous baseline with more benign data, but also with mixed device profiles. Therefore, the per-device/global comparison changes both specialization and training-data composition/volume (Table 3). Differences in performance,

therefore, reflect the combined effect of training-data composition and anomaly-scoring behavior, which is evaluated explicitly in the ablation reported in Section 4.1.

3.5. Reconstruction Scores and Threshold Calibration

For each scaled flow x, the scalar score supplied to EVT and to the threshold-independent metrics is the sample-level root mean squared reconstruction error (RMSE):

$$\mathrm{RMSE}(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}. \tag{1}$$

EVT operating point. The threshold is calibrated from benign validation RMSE values. The baseline is the 90th percentile, corresponding to tail fraction $q = 0.10$. A GPD with location fixed at zero is fitted to non-negative exceedances, and its conditional quantile targets nominal aggregate FPR $\alpha = 0.01$. When fewer than ten exceedances are available, the implementation falls back to the empirical $Q_{1-\alpha}$ quantile. The code flags equality at the threshold:

$$A_{\mathrm{EVT}}(x) = \mathbb{I}[\mathrm{RMSE}(x) \geq \tau_{\mathrm{EVT}}]\,. \tag{2}$$

The target FPR is a calibration objective, not a guarantee; empirical FPR is measured on independent benign test flows.

LRE operating point. For feature i, LRE uses the absolute reconstruction deviation $e_i(x) = |x_i - \hat{x}_i|$. A separate GPD threshold is calibrated from each benign validation-error distribution using $q = 0.10$ and nominal per-feature FPR $\alpha_f = 0.005$. The empirical $Q_{1-\alpha_f}$ quantile is used when fewer than ten exceedances are available or the GPD fit fails. The decision is

$$A_{\mathrm{LRE}}(x) = \mathbb{I}[\exists i :\ |x_i - \hat{x}_i| > \tau_i]\,. \tag{3}$$

As LRE applies a logical union across 28 feature-level decisions, its aggregate flow-level FPR can exceed the nominal per-feature target. EVT and LRE are treated as two operating points: EVT prioritizes a lower alert rate, whereas LRE prioritizes sensitivity to localized deviations. Supporting ablation also derives OR and AND combinations from the two binary flag sets; these combinations do not introduce separately calibrated thresholds.

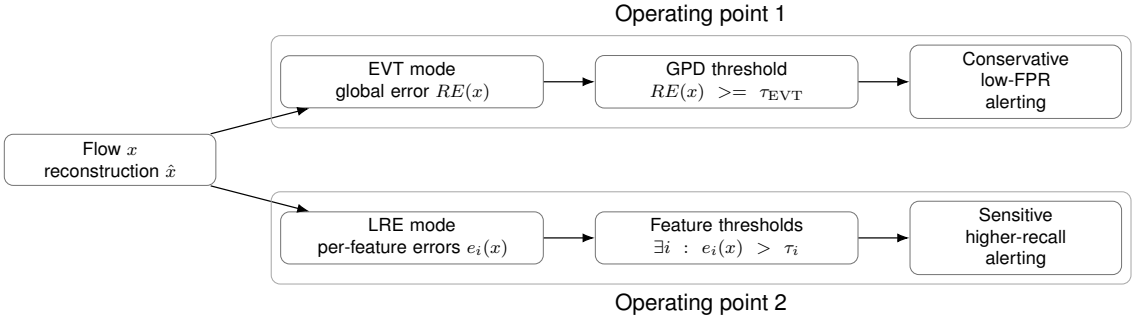

Figure 2. Two scoring modes evaluated in the framework. EVT uses a global reconstruction-error threshold, whereas LRE applies per-feature thresholds and an OR rule.

3.6. Evaluation and Statistical Analysis

Threshold-dependent evaluation reports Detection Rate (DR), FPR, precision, F1-score, specificity, and confusion-matrix counts. AUROC is computed before thresholding from the continuous sample-level RMSE scores; it therefore characterizes scalar-score ranking and is not an AUROC of the discrete EVT or LRE decisions. The implementation also calculates Average Precision from the same RMSE scores, but it is not used in the main analysis.

Empirical FPR uncertainty is summarized with 95% Wilson binomial intervals. Repeated-holdout comparisons use one (device, attack, repetition) tuple as the paired analytical unit, yielding $4 \times 18 \times 5 = 360$ observations. Per-subgroup differences between per-device and global models are assessed with paired Wilcoxon signed-rank tests, with Holm adjustment across the four subgroup tests. The table reports the resulting adjusted p-values. Cohen's d_z is reported descriptively and is computed as the mean paired difference divided by the standard deviation of the paired differences. The 95% bootstrap confidence intervals for mean and median paired differences use 10,000 resamples. Since the repeated partitions overlap and observations share devices and attacks, these tests are interpreted as stability and sensitivity analyses, not as inference from fully independent replications.

3.7. Feature Attribution and Execution-Cost Measurement

SHAP is applied after detection to absolute per-feature reconstruction errors for flows flagged by the per-device LRE models. Mean absolute SHAP values rank the features associated with each anomaly score. The analysis is restricted to behavioral attacks, for which the largest per-device/global differences are observed. The attributions are associative and do not establish causal relationships.

Execution cost is measured on persisted 28-feature models using Python 3.11.9, PyTorch 2.11.0 with CUDA 12.6, and an NVIDIA GeForce RTX 4060 Laptop GPU. Single-flow latency uses 100 warmup calls followed by 1,000 timed executions. Batch throughput uses batches of 256 flows, five warmup batches, and 20 timed repetitions. Synthetic Gaussian inputs matching the persisted feature dimensionality are used, so the measurements include model and scoring execution but exclude dataset loading, feature extraction, network I/O, device identification, and end-to-end gateway latency. Model size includes the persisted scaler, autoencoder, thresholds, and metadata. Training time, energy consumption, SHAP latency, and model-update cost were not recorded.

4. Evaluation, Results, and Discussion

This section reports the empirical results according to the research questions defined in Section 3.1 and discusses their security, statistical, and operational implications.

4.1. RQ1: Per-Device Versus Global Detection

Table 4 summarizes the like-for-like paired comparison between per-device models and the global baseline with LRE scoring held fixed, aggregated by attack subgroup across the five repeated holdout partitions. The unit of analysis is one (device, attack, repetition) tuple. "Per-device mean DR" and "Global mean DR" are arithmetic means over those units; "Mean pairwise Δ" is the mean of the per-unit differences (Per-device $-$ Global),

Table 4. Detection rate under matched LRE scoring: per-device vs. global baseline. Unit: one (device, attack, repetition) tuple (n paired observations per subgroup). The mean pairwise delta equals the difference of paired means up to rounding; the median is computed unit-by-unit.

Attack Subgroup	n	Per-Dev mean DR	Global mean DR	Mean pair. Δ	Median pair. Δ	95% CI of Δ	Cohen's d_z	Holm-adjusted p
Behavioral Protocol Attacks	120	0.949	0.682	+0.267	+0.230	[+0.235, +0.301]	1.46	9.85×10^{-22}
MQTT Connect Floods	40	1.000	0.855	+0.145	+0.113	[+0.116, +0.174]	1.53	1.75×10^{-8}
MQTT Publish Floods	40	0.999	0.907	+0.093	+0.051	[+0.066, +0.123]	1.01	1.77×10^{-8}
Volumetric Network Floods	160	1.000	0.936	+0.064	+0.000	[+0.036, +0.095]	0.34	1.19×10^{-25}

which equals the difference of the paired marginal means up to displayed rounding; "95% CI" is a 10,000-resample bootstrap confidence interval on the mean pairwise delta.

All four paired Wilcoxon tests reject the null after Bonferroni–Holm correction. Wins/losses/ties at the unit level are 120/0/0 for behavioral attacks, 40/0/0 for MQTT connect floods, 40/0/0 for MQTT publish floods, and 32/0/128 for volumetric floods. Volumetric attacks require a careful reading: the median pairwise delta is zero, 128 of 160 units are exact ties, and the positive mean is driven by the remaining 32 units. Accordingly, the small effect size for volumetric floods ($d_z = 0.34$) is more informative operationally than the very small p-value. The behavioral effect is substantially larger ($d_z = 1.46$).

The most pronounced advantage occurs for behavioral protocol attacks. Under the same LRE scoring mode, the global baseline achieves a mean DR of 0.682, compared with 0.949 for per-device models across the repeated holdouts. Reconnaissance, ARP spoofing, and malformed MQTT data produce less saturated detection rates than most volumetric floods, leaving greater scope for model composition to affect the learned normality boundary. The current evidence is consistent with device-dependent score distributions, but the mixed composition of the global pool prevents the gain from being attributed exclusively to specialization.

Table 5 provides the threshold-independent AUROC results computed from sample-level RMSE on the main test partition. Per-device AUROC is averaged over the device–attack pairs in each subgroup; global AUROC is averaged over the corresponding attack types. The largest difference again occurs for behavioral attacks (0.982 vs. 0.891), whereas the volumetric difference is small (0.998 vs. 0.991).

To separate the contribution of *specialization* (per-device vs. global training) from the contribution of *scoring mode* (EVT vs. LRE), Table 6 reports a four-way ablation with the same architecture, hyperparameters, and EVT calibration procedure for all four configurations.

Reading Table 6 along the rows separates the two design choices at the reported operating points. (i) *Training composition*: holding the scoring mode fixed, the per-device configuration improves DR over the mixed global baseline in all subgroups under both EVT (e.g., 0.836 vs. 0.524 for behavioral attacks) and LRE (0.945 vs. 0.681). In the

Table 5. AUROC computed from continuous sample-level RMSE by attack subgroup on the main test partition.

Subgroup	Per-Device AUROC	Global AUROC	Difference
Behavioral Protocol Attacks	0.982	0.891	+0.091
MQTT Connect Floods	0.996	0.966	+0.030
MQTT Publish Floods	0.997	0.964	+0.033
Volumetric Network Floods	0.998	0.991	+0.007

Table 6. Ablation on the main test partition: detection rate / aggregate FPR / F1 by training composition (per-device vs. global) and scoring mode (EVT vs. LRE). Means by attack subgroup.

Subgroup	Per-Device EVT DR / FPR / F1	Per-Device LRE DR / FPR / F1	Global EVT DR / FPR / F1	Global LRE DR / FPR / F1
Behavioral Protocol Attacks	0.836 / 0.012 / 0.892	0.945 / 0.063 / 0.952	0.524 / 0.009 / 0.617	0.681 / 0.090 / 0.633
MQTT Connect Floods	0.911 / 0.012 / 0.945	1.000 / 0.063 / 0.998	0.103 / 0.009 / 0.185	0.869 / 0.090 / 0.896
MQTT Publish Floods	0.948 / 0.012 / 0.970	1.000 / 0.063 / 0.998	0.495 / 0.009 / 0.594	0.862 / 0.090 / 0.900
Volumetric Network Floods	0.942 / 0.012 / 0.946	1.000 / 0.063 / 1.000	0.749 / 0.009 / 0.752	1.000 / 0.090 / 0.998

absence of a cameras-only pooled control, these differences reflect per-device versus mixed global composition rather than specialization alone. (ii) *Scoring mode*: holding training composition fixed, switching from EVT to LRE increases DR but also raises aggregate FPR from approximately 1.2% to 6.3% for per-device models, and from approximately 0.9% to 9.0% for the global baseline. The combination "per-device + LRE" yields the highest mean F1 across all four subgroups in this evaluation, whereas "per-device + EVT" provides the lower-FPR per-device point. Both are presented as distinct operating points, with no single configuration recommended universally.

4.2. RQ2: EVT and LRE as Operating Points

For per-device behavioral attacks, EVT yields mean DR/FPR/precision/F1 of 0.836/0.012/0.987/0.892, whereas LRE yields 0.945/0.063/0.964/0.952. These values are already represented in the four-way ablation and define the lower-alert and higher-sensitivity operating points, respectively.

EVT mode achieves an aggregate FPR of approximately 1.2% on benign test traffic and a mean precision of 0.99; LRE mode achieves a higher behavioral detection rate (0.945) at the cost of an aggregate FPR of approximately 6.3%. LRE detects nearly all evaluated volumetric flows, whereas the per-device EVT mean is 0.942 and remains lower for the UDP variants shown in the subgroup results. The choice of scoring mode has its strongest effect on behavioral and MQTT attacks, although some volumetric subtypes are also affected. The analysis does not claim that one mode is universally better; the two modes are presented as configurable operating points whose suitability depends on the alert volume that operators are willing to inspect.

The benign-test FPR ablation further shows that, for every evaluated model, the FPR of the EVT OR LRE rule is identical to the LRE-only FPR, and EVT AND LRE is identical to EVT-only FPR. Under these calibrations, EVT alerts are contained within the more sensitive LRE alert set on benign traffic. The OR and AND rules do not create additional empirical FPR operating points in the reported configuration.

EVT is treated as a statistically motivated threshold not as a calibration that strictly controls FPR. Under the assumptions of the Pickands-Balkema-de Haan theorem, the GPD threshold targets a 1% FPR; the empirical aggregate FPRs measured on benign test traffic, 0.95–1.42% per device, are near the nominal target but not exactly 1%, and depend on the chosen tail fraction ($q = 0.10$ in this study). A systematic sensitivity analysis of q, the GPD goodness-of-fit, and alternative thresholds (percentile, MAD, Gaussian) is left for future work.

4.3. RQ3: SHAP Feature Attribution

Table 7 reports the top-three features for ARP Spoofing as a representative case, measured by mean absolute SHAP value over test samples flagged by the per-device LRE detector. The subsequent frequency analysis covers ARP Spoofing, MQTT-Malformed Data, and Recon-VulScan.

Table 7. Top-three SHAP features by device for ARP Spoofing (mean absolute SHAP value).

Rank	Owltron Camera	Ecobee Camera	M1T Camera	Blink Camera
1	Max (0.033)	Variance (0.031)	Srate (0.038)	Protocol Type (0.358)
2	Variance (0.032)	Header_Length (0.029)	Header_Length (0.037)	Duration (0.085)
3	rst_count (0.025)	Duration (0.025)	rst_count (0.030)	psh_flag_number (0.048)

Frequency of Variance in the top-3. Across the three behavioral attacks analyzed, Variance appears in the top-3 SHAP features for $3/3$ device–attack pairs on Owltron, $2/3$ on Ecobee, $2/3$ on M1T, and $1/3$ on Blink, for a total of $8/12$ pairs. Variance is frequently relevant, but it is not universal and occupies rank 1 in only $2/12$ pairs. The attribution evidence supports recurrent, device-dependent feature rankings rather than a single dominant feature.

Beyond Variance, statistical flow features (Header_Length, Duration, Max, Srate) and TCP control-flag counters (rst_count, syn_count) appear repeatedly across devices, showing that the anomaly scores draw on a recurring but non-identical set of flow descriptors. The exact top-ranked composition varies by device; for example, Protocol Type has the largest mean absolute SHAP value for Blink ARP Spoofing, and Max, Variance, and Srate lead for other devices. Table 7 does not include benign feature distributions or causal interventions, so explanations based on a "narrow" or "more variable" normal profile remain hypotheses, not demonstrated mechanisms. The defensible conclusion is limited to device-dependent attribution rankings.

4.4. RQ4: Repeated-Holdout Stability

Table 8 reports the per-device LRE detection rate for behavioral protocol attacks aggregated over five repeated random holdout partitions. The mean and standard deviation are computed over the $5 \times 6 = 30$ (attack, repetition) units per device. The partitions are generated with different seeds and may overlap; they are not classical mutually exclusive k-fold partitions.

Table 8. Detection-rate stability across five repeated random holdout partitions (per-device LRE, behavioral protocol attacks). Mean / std / min / max are computed over the 30 (attack, repetition) units per device.

Device	Mean DR	Std DR	Min DR	Max DR
Blink Camera	0.997	0.004	0.988	1.000
M1T Camera	0.982	0.023	0.942	1.000
Owltron Camera	0.914	0.113	0.674	1.000
Ecobee Camera	0.902	0.121	0.629	1.000

The repeated-holdout summary is stable for Blink and M1T but more variable for Owltron and Ecobee. The minimum DR observed is 0.629 (Ecobee Camera, ARP Spoofing, one repetition), followed by 0.674 for Owltron ARP Spoofing and 0.942 for M1T ARP Spoofing. Attack-specific stability results identify ARP Spoofing as the largest source of variability for Owltron and Ecobee: its mean DR is approximately 0.719 ($\mathrm{SD} = 0.041$) and 0.665 ($\mathrm{SD} = 0.031$), respectively. Recon-VulScan and MQTT-Malformed Data also vary for some devices, but to a lesser degree. The available results show sensitivity to split selection without isolating whether the cause is attack-sample composition, benign-boundary variation, or another partition-dependent factor.

The pooled repeated-holdout Wilcoxon comparisons against the global baseline remain significant after Bonferroni–Holm correction for all four subgroups. The direction of the matched-LRE difference is consistent across the reported units, but the overlapping partitions and shared devices/attacks mean that the very small p-values should not be read as evidence from independent replications.

4.5. Security Interpretation of the Results

The largest observed gains for per-device models occur on behavioral protocol attacks. Under matched LRE scoring, the global baseline reaches a mean DR of 0.682 against 0.949 for the per-device models across the repeated holdouts (Table 4), a mean pairwise gain of $+0.267$ with a 95% bootstrap CI of $[+0.235, +0.301]$. The main-split ablation in Table 6 shows the same direction under both scoring modes: with EVT fixed, behavioral DR increases from 0.524 (global) to 0.836 (per-device), and with LRE fixed, it increases from 0.681 to 0.945. The AUROC comparison is also consistent with improved score separation for the per-device models (0.982 vs. 0.891). These results do not depend on a cross-mode comparison, yet the mixed composition of the global pool remains a confounder.

A plausible interpretation is that a device-specific model preserves local reconstruction boundaries that can be blurred by a mixed training pool. The SHAP results are compatible with this interpretation as the leading reconstructed features vary across devices, but they do not establish the underlying benign-profile mechanism. In particular,

the large Protocol Type attribution for Blink and the higher split sensitivity of Ecobee ARP Spoofing are observations, not proof that one device has a narrower or more variable normal profile. Establishing such explanations would require direct analysis of benign feature distributions and reconstruction-error separation.

Volumetric attacks tell a different story. Under repeated matched-LRE comparisons, the median pairwise delta is zero, 128 of 160 units tie, and the effect size is small ($d_z = 0.34$). On the main partition, the LRE means are both approximately 1.000, with a difference of only about $+0.0004$; the larger EVT difference (0.942 vs. 0.749) is concentrated in the conservative operating point. Per-device modeling is therefore not interpreted as necessary for volumetric detection, and the matched-LRE benefit is operationally negligible despite statistical significance. Overall, the results consistently indicate that device-specific benign modeling improves sensitivity to behavioral anomalies under controlled offline conditions in CICIoMT2024.

4.6. Deployment Costs and Operating Context

Table 9 reports the execution-cost measurements described in Section 3.7. These measurements characterize individual persisted models on one workstation GPU and do not represent end-to-end gateway or on-device deployment. EVT is the lower-alert point, whereas LRE is the higher-sensitivity point; the appropriate choice depends on the alert volume that operators can inspect.

Table 9. Measured model-execution cost. Latency is mean / p95 for one 28-feature input; throughput uses batch size 256.

Model	EVT Lat. (ms)	EVT Thr. (flows/s)	LRE Lat. (ms)	LRE Thr. (flows/s)	Disk (MB)
Owltron Camera	0.565 / 0.741	421,673	0.540 / 0.709	399,953	1.154
Ecobee Camera	0.602 / 0.933	418,434	0.560 / 0.781	440,358	1.084
M1T Camera	0.569 / 0.777	427,243	0.547 / 0.716	421,389	1.041
Blink Camera	0.564 / 0.774	419,810	0.547 / 0.738	436,090	1.030
Global Baseline	0.558 / 0.736	419,407	0.548 / 0.746	405,992	2.224

Single-flow mean latency is 0.54–0.60 ms, p95 remains below 0.94 ms, and batch throughput is approximately 400,000–440,000 flows/s. The four per-device models occupy 4.31 MB versus 2.22 MB globally ($1.94\times$). Per-request latency does not scale with model count when routing queries only the associated model, but storage and resident memory do. Training time, energy, SHAP overhead, update cadence, simultaneous residency, preprocessing, routing, and network I/O are unmeasured. As the benchmark used an RTX 4060 Laptop GPU, it does not establish ESP32-class or hospital-gateway readiness [Cavalcante et al. 2024].

5. Threats to Validity

All experiments use CICIoMT2024 and only four camera-class IoMT devices, so the findings should not be generalized to other device classes or deployments without further evaluation. The choice of cameras enables a controlled within-class study but does not establish that the same gains occur for wearables, pumps, monitors, or broader IoT and cyber-physical systems. The global baseline contains mixed device profiles beyond the

four evaluated camera partitions and no cameras-only pooled control is reported; hence, per-device/global differences do not isolate specialization from training composition and sample volume.

The use of random flow-level splits, rather than temporal or per-capture splits, may underestimate detection difficulty under realistic temporal variation, near-duplicate flows, or benign drift. The five repeated holdout partitions overlap and share devices and attacks, so their paired units are not independent replications and the magnitudes of very small p-values should not be overinterpreted. Offline routing uses the dataset's ground-truth device label. MAC/IP identification, authenticated device binding, identity spoofing, and flow misrouting are outside the experiment; ARP Spoofing detection should therefore not be interpreted as evidence that the model-selection mechanism is robust to spoofed identity.

The evaluated autoencoder is overcomplete and no compressive-bottleneck variant, architecture ablation, or direct benign-versus-attack reconstruction-gap analysis is reported. EVT calibration uses a fixed configuration ($q = 0.10$, $\alpha = 0.01$), without GPD goodness-of-fit analysis, sensitivity evaluation of the tail fraction q, or attack-performance comparison against percentile, MAD, or Gaussian thresholds. The comparison is limited to per-device and global autoencoders with EVT/LRE scoring and does not include other one-class detectors or a supervised reference model. SHAP rankings are associative and do not establish causal explanations of device behavior.

The framework detects anomalous flows but does not classify attack types, and application-layer payload behavior is outside the scope of network-flow detection. Execution measurements cover single-model latency, batch throughput, and on-disk size on one RTX 4060 Laptop GPU using synthetic inputs. They exclude feature extraction, network I/O, online device identification, end-to-end gateway latency, training time, energy, SHAP overhead, simultaneous multi-model residency, and model-update cost.

6. Conclusion and Future Work

This paper compared per-device and mixed-global benign-only autoencoders for four CIC-IoMT2024 cameras under matched EVT and LRE scoring. On the main split, behavioral DR increases from 0.524 to 0.836 under EVT and from 0.681 to 0.945 under LRE; AUROC increases from 0.891 to 0.982. The repeated-holdout matched-LRE gain is $+0.267$ (95% CI $[+0.235, +0.301]$, $d_z = 1.46$). For volumetric attacks, LRE is close to the empirical ceiling and the matched benefit is operationally negligible despite statistical significance.

EVT provides the lower-alert point and LRE the higher-sensitivity point. SHAP shows recurrent but device-dependent feature rankings, not a universal explanatory feature. Measured execution is below 0.61 ms mean latency with approximately 400–440 thousand flows/s; maintaining four models increases disk storage from 2.22 to 4.31 MB. These results support per-device screening principally for behavioral deviations, subject to correct routing, sufficient benign data, and the mixed-baseline limitation; they do not establish end-to-end deployment readiness.

Future work should evaluate additional device classes and datasets, a cameras-only pooled control, other one-class and supervised references, matched-FPR and simpler-threshold comparisons, compressive architectures, temporal/capture splits and controlled

drift, spoofing-resistant routing, ransomware-specific traffic, and end-to-end gateway, energy, SHAP, and update costs.

Acknowledgments

This work has been fully funded by the projects *Xeque-Mate Agente & Agregador* supported by Centro Integrado de Segurança em Sistemas Avançados (CISSA), with financial resources from the PPI IoT/Manufatura 4.0 of the MCTI grant number CIS-AFCCT-2025-5-1-2 & CIS-AFCCT-2025-5-1-1, signed with EMBRAPII.

References

Cavalcante, J., Barros, T. G. F., and de Souza, J. N. (2024). Autonomous network intrusion detection for resource-constrained devices of the internet of things. In *Anais do XXIV Simposio Brasileiro de Seguranca da Informacao e de Sistemas Computacionais*. SBC.

Dadkhah, S., Pinto Neto, E. C., Ferreira, R., Molokwu, R. C., Sadeghi, S., and Ghorbani, A. A. (2024). CICIoMT2024: A benchmark dataset for multi-protocol security assessment in IoMT. *Internet of Things*, 25:101080.

Damasceno, M. G. L., Souza, C. B. B., and Balieiro, A. M. (2025). Evaluating MLP and autoencoder models for zero-day attack detection in 6G networks. In *Anais do XXV Simposio Brasileiro de Seguranca da Informacao e de Sistemas Computacionais*. SBC.

Emelianova, N., Kamienski, C., and Prati, R. C. (2025). Optimizing IoT threat detection with kolmogorov-arnold networks (KANs). In *Anais do XXV Simposio Brasileiro de Seguranca da Informacao e de Sistemas Computacionais*. SBC.

Gorman, M., Ding, X., Maguire, L., and Coyle, D. (2023). Anomaly detection in batch manufacturing processes using localized reconstruction errors from 1-d convolutional autoencoders. *IEEE Transactions on Semiconductor Manufacturing*, 36(1):147–150.

Hernandez-Jaimes, M. L., Martinez-Cruz, A., Ramirez-Gutierrez, K. A., and Guevara-Martinez, E. (2024). Enhancing machine learning approach based on nilsimsa fingerprinting for ransomware detection in IoMT. *IEEE Access*.

Kavkas, E. and Yildiz, K. (2025). Enhancing IoMT security with deep learning based approach for medical IoT threat detection. In *Proceedings of the IEEE International Symposium on Digital Forensic and Security*.

Khraisat, A., Gondal, I., Vamplew, P., and Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity*, 2(1):20.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Quincozes, C. B., Oliveira, H. C., Quincozes, S. E., Miani, R. S., and Quincozes, V. E. (2024). Uma arquitetura baseada em inteligencia artificial explicavel (XAI) para sistemas de deteccao de intrusoes em smart grids. In *Anais do XXIV Simposio Brasileiro de Seguranca da Informacao e de Sistemas Computacionais*. SBC.

Ramesh, K., Miller, N. C., and Faridi, A. (2024). Efficient machine learning frameworks for strengthening cybersecurity in internet of medical things (IoMT) ecosystems. *arXiv preprint arXiv:2412.01375*. Preprint; venue to be confirmed before camera-ready.

Sakurada, M. and Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, MLSDA'14, page 4–11, New York, NY, USA. Association for Computing Machinery.

Siffer, A., Fouque, P.-A., Termier, A., and Largouet, C. (2017). Anomaly detection in streams with extreme value theory. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Simioni, J. A., Viegas, E. K., Santin, A. O., and Horchulhack, P. (2024). Deteccao de intrusao atraves de redes neurais profundas com saidas antecipadas para inferencia rapida e confiavel. In *Anais do XXIV Simposio Brasileiro de Seguranca da Informacao e de Sistemas Computacionais*. SBC.

Sivanathan, A., Gharakheili, H. H., Loi, F., Radford, A., Wijenayake, C., Vishwanath, A., and Sivaraman, V. (2019). Classifying iot devices in smart environments using network traffic characteristics. *IEEE Transactions on Mobile Computing*, 18(8):1745–1759.

Sohail, F., Bhatti, M. A. M., Awais, M., and Iqtidar, A. (2024). Explainable boosting ensemble methods for intrusion detection in internet of medical things (IoMT) applications. *IEEE Access*. Venue confirmed as IEEE Access; DOI to be added in camera-ready.

Vieira, L. Q., Choren, R., and Sant'ana, R. (2025). Contrastive autoencoding with gaussian confidence regions for concept drift detection in IDS. In *Anais do XXV Simposio Brasileiro de Seguranca da Informacao e de Sistemas Computacionais*. SBC.

Yacoubi, M., Moussaoui, O., and Drocourt, C. (2026). AI for IoMT security: A comprehensive survey of intrusion detection and system architectures. *Internet of Things*, 36:101869.