Beyond Aggregate Accuracy: Kill Chain-Aware Evaluation of IoMT Intrusion Detection Models

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], José A. Suruagy [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. *Machine learning-based Intrusion Detection Systems (IDSs) have shown promising results for Internet of Medical Things (IoMT) environments; however, aggregate performance metrics can conceal failures in detecting early-stage behaviors. This paper presents a kill chain-aware analysis of Random Forest, LightGBM, a 1D-CNN, and an autoencoder using the CICIoMT2024 dataset, analyzing detection capability across 18 attack categories and multiple stages of the attack lifecycle. Results show that Random Forest substantially outperforms the CNN in reconnaissance and spoofing scenarios, while default LightGBM configurations exhibit severe degradation when transitioning from binary to multi-class classification. In addition, the benign-only autoencoder achieves high attack recall without requiring attack labels, although its false-positive burden limits direct alert generation. Supported by bootstrap confidence intervals and McNemar's test, the findings demonstrate that IoMT IDSs should move beyond aggregate accuracy toward phase-aware, class-disaggregated, and statistically grounded assessment protocols.*

1. Introduction

Modern hospital infrastructures depend on interconnected IoMT devices, including infusion pumps, patient monitors, and imaging systems, that commonly communicate through Wi-Fi and MQTT and are managed at centralized gateways. This connectivity expands the attack surface, but also makes gateways practical IDS observation points for intercepting malicious traffic before lateral propagation. WannaCry [Ehrenfeld 2017] and later ransomware-as-a-service operations [Ruellan et al. 2024] illustrate the operational consequences of healthcare cyber attacks.

Recent studies on ML-based IDS for IoMT frequently report binary accuracy above 99% [Dadkhah et al. 2024, Sohail et al. 2024, Mohammadi et al. 2024, Kavkas and Yildiz 2025, Akar et al. 2025]. However, severe class imbalance allows high-volume attacks to dominate aggregate metrics, potentially hiding failures in less frequent

reconnaissance and spoofing behaviors that are relevant to preventive defense. High binary accuracy may separate benign from malicious traffic without establishing whether early-stage, rare, or clinically consequential attacks are detected. Accordingly, this paper tests the hypothesis that aggregate IDS metrics are insufficient for assessing operational defensive capability, as high binary accuracy may separate benign from malicious traffic, but it does not reveal whether early-stage, rare, or clinically consequential attacks are detected.

The evaluation addresses three questions: Does detection capability remain consistent across attack phases as classification granularity increases from binary to 19-class detection? To what extent can default model configurations lead to misleading perceptions of defensive capability? Can a benign-only autoencoder provide useful anomaly screening as a complement to supervised attack classification? The analysis uses accessible baseline architectures on CICIoMT2024 [Dadkhah et al. 2024] to quantify defensive capability and operational limitations under configurations that do not require extensive architecture engineering.

The contributions are: (i) phase- and class-level evidence comparing the evaluated 1D-CNN with tree-based baselines on early-stage attacks; (ii) quantification of degradation caused by the default LightGBM configuration; (iii) evaluation of benign-only anomaly screening; and (iv) SHAP/LIME inspection to determine whether influential features correspond to plausible network indicators instead of obvious dataset artifacts. Section 2 reviews the context; Section 3 defines the protocol; Sections 4-5 report results and validation; and Sections 6-7 discuss limitations and deployment implications.

2. Background and Related Work

The cyber kill chain [Hutchins et al. 2011] models intrusion progression, while MITRE ATT&CK [MITRE ATT&CK 2025] structures adversarial tactics and techniques. At IoMT gateways, network flows may expose malicious behavior before service disruption. Accordingly, CICIoMT2024 classes are mapped to *reconnaissance* (host, port, OS, and vulnerability enumeration), *access and exploitation* (spoofing and protocol misuse), and *impact* (DoS, DDoS, and MQTT abuse). This mapping supports defender-oriented interpretation of per-class results without asserting a strictly linear sequence (Figure 1).

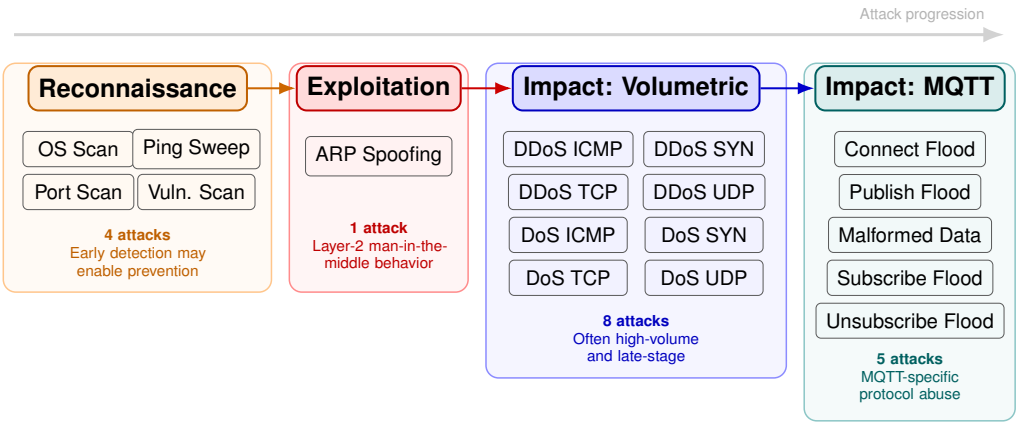

Figure 1. CICIoMT2024 classes mapped to defender-relevant phases; the mapping does not imply a strictly linear attack sequence.

Table 1. Class distribution in the CICIoMT2024 test partition (derived from the evaluated test files; total = 1,614,182 samples).

Category	Class	Samples	Category	Class	Samples
DDoS	DDoS-UDP	362,070	DoS	DoS-ICMP	98,432
DDoS	DDoS-ICMP	349,699	DoS	DoS-SYN	98,595
DDoS	DDoS-TCP	182,598	DoS	DoS-UDP	137,553
DDoS	DDoS-SYN	172,397	DoS	DoS-TCP	82,096
Benign	Benign	37,607	Spoofing	ARP Spoofing	1,744
MQTT	DDoS-Connect Flood	41,916	Recon	Port Scan	22,622
MQTT	DDoS-Publish Flood	8,416	Recon	OS Scan	3,834
MQTT	DoS-Publish Flood	8,505	Recon	VulScan	1,034
MQTT	DoS-Connect Flood	3,131	Recon	Ping Sweep	186
MQTT	Malformed Data	1,747			

Gateway IDSs convert flow-level evidence into detection decisions. Supervised models learn labeled mappings to known attack classes; Random Forest [Breiman 2001] and LightGBM [Ke et al. 2017] are strong candidates for tabular traffic, while 1D-CNNs are widely used deep-learning representatives [Vinayakumar et al. 2019]. Benign-only autoencoders instead learn normal traffic and flag reconstruction deviations [Goodfellow et al. 2016]. The two settings serve complementary roles: supervised models provide class-specific attribution, whereas anomaly detectors can screen deviations when attack labels are incomplete or delayed.

CICIoMT2024 studies report high aggregate results in binary, coarse-grained, or 19-class settings [Dadkhah et al. 2024, Sohail et al. 2024, Mohammadi et al. 2024, Kavkas and Yildiz 2025, Akar et al. 2025], but generally do not focus on phase-level behavior across reconnaissance, spoofing, and MQTT attacks. Chandekar et al. [Chandekar et al. 2025] and Domenech et al. [Doménech et al. 2025] add statistical validation; the latter also reports cross-dataset degradation, indicating that point estimates can overstate deployment readiness. Yacoubi et al. [Yacoubi et al. 2026] apply explainability in the binary setting. In adjacent contexts, Damasceno et al. [Damasceno et al. 2025] observe substantial attack-dependent autoencoder variation and argue for hybrid use, while one-class or hierarchical gateway detection is investigated in [Ayad et al. 2025, Adji et al. 2025, Uddin et al. 2025]. The question of whether aggregate results translate into consistent behavior across defender-relevant attack phases remains less explored. Collectively, prior works motivate evaluating IoMT intrusion detection combining phase-aware analysis, benign-only anomaly screening, operational profiling, feature-attribution inspection, and statistical validation.

3. Methodology

3.1. Dataset and Scope

Experiments use the publicly available processed Wi-Fi/MQTT feature-level CICIoMT2024 data [Dadkhah et al. 2024]: 7,160,831 training and 1,614,182 test flows, totaling approximately 8.8 million records. BLE is outside the main scope and discussed in Section 6. The original dataset test distribution (Table 1) is evaluated at three levels of granularity: binary detection (Benign versus Attack), six-class classification (Benign, DDoS, DoS, MQTT, Recon, and Spoofing), and 19-class classification covering all indi-

Table 2. Feature taxonomy of the 45 flow-level features used in this study, grouped by functional category and security relevance.

Category	Count	Security relevance
Temporal indicators	5	Duration, Rate, Srate, Drate, IAT — discriminate flooding (low IAT, high rate) from probing (sparse timing).
TCP flags	7	SYN, ACK, RST, FIN, PSH, ECE, CWR — expose handshake anomalies, SYN floods, incomplete teardowns.
TCP counters	4	Cumulative flag counts — aggregate connection-level evidence over the flow.
Protocol indicators	15	Binary presence flags (ARP, ICMP, IGMP, IPv4, MQTT, DNS, HTTP, HTTPS, etc.) — identify Layer-2/3/7 abuse vectors.
Packet statistics	4	Number, Tot_size, Min, Max — distinguish probes (sparse, small) from floods (dense, uniform).
Flow descriptors	6	Header_Length, AVG, Std, Magnitude, Variance, Weight — aggregate flow shape and dispersion.
Volume / type	4	Total volume, packet length variance, protocol type indicators.
Total	**45**	—

vidual attack types. These settings correspond to progressively richer operational outputs, from rapid alerting to attack-family triage and fine-grained support for incident response and forensic analysis.

3.2. Preprocessing

The preprocessing pipeline ensures consistent feature representation while preventing information leakage across data splits. Train and test partitions provided by the dataset authors [Dadkhah et al. 2024] are loaded separately, retaining only numeric columns cast to `float32`. Labels are encoded and remapped to the target classification scenario (2, 6, or 19 classes), and the training partition is then split into 80% train and 20% stratified validation, preserving class proportions. An integrity audit over both the training and test partitions confirmed zero missing values in the evaluated processed files; the mean-imputation step was therefore retained only as a defensive preprocessing safeguard and is fitted with per-column means computed exclusively on the training subset; features are subsequently z-score normalized using a `StandardScaler` fitted on the training subset only and then applied identically to the validation and test partitions. For the convolutional model, the feature vector is reshaped to $(N, 45, 1)$.

Due to their optimization procedures, different model families consume distinct effective training volumes. Random Forest and Logistic Regression do not require validation-based early stopping and therefore train on the combined train and validation subsets ($\sim$7.16M samples). LightGBM (in its tuned variants) and the 1D-CNN keep the validation split separate to support early stopping ($\sim$5.73M + $\sim$1.43M samples). The autoencoder is trained exclusively on benign samples ($\sim$165,000) drawn from the pooled train and validation partitions, since labeled attack data is by design unavailable to the anomaly screening setting; this asymmetry is acknowledged as a threat to validity in Section 6.

The test set preserves the original CICIoMT2024 class distribution, including the low prevalence of reconnaissance samples, so that models are evaluated under the dataset's unbalanced class priors. Each flow is represented by 45 numerical features derived from the processed CICIoMT2024 traffic representation [Dadkhah et al. 2024] and organized in

Table 3. Hyperparameter summary for the evaluated IDS components. "ES" denotes early stopping patience.

Component	Key hyperparameters	Class balancing
Random Forest	100 trees, $\sqrt{d}$=6 feat./split, depth ∞, $n_{\min}$=2, bootstrap	None
Logistic Reg.	L2 (C=1.0), L-BFGS, 1,000 iter, multinomial	None
LightGBM (default)	100 est., η=0.1, 31 leaves, no ES	None
LightGBM (tuned)	500–800 est., η=0.01–0.05, 63–127 leaves, ES = 30, variant select.	`is_unbalance`
1D-CNN	$\text{Conv}_{32}\rightarrow\text{Pool}_2\rightarrow\text{Conv}_{64}\rightarrow\text{Pool}_2\rightarrow\text{Dense}_{128}$, Adam η=10^{-3}, ES = 3	Inv.-freq. weights
Autoencoder	Enc. $45\rightarrow64\rightarrow32\rightarrow16$, BN, Dropout = 0.2, MSE, ES = 5	Benign-only

this study into security-relevant groups (Table 2); all features are retained so that every model is evaluated on the same input space.

3.3. Detection Settings

Two complementary gateway settings are evaluated independently: *Anomaly Screening*. An autoencoder trained exclusively on benign traffic computes reconstruction error for each incoming flow. Flows whose error exceeds a calibrated threshold trigger anomaly alerts, enabling detection of traffic patterns that deviate from learned normality without requiring labeled attack data. *Attack Classification.* Supervised classifiers categorize flows into 18 attack types mapped to MITRE ATT&CK-informed phases (Section 2), providing the specific threat identification needed for incident triage.

3.4. Model Selection Rationale

The study uses standard baseline models to assess deployable IDS capability rather than introduce new architectures. Tree-based methods serve as strong references for tabular data [Grinsztajn et al. 2022]. The analysis examines whether aggregate performance remains consistent across security-relevant phases, particularly reconnaissance and spoofing versus volumetric denial-of-service attacks. Table 3 summarizes the hyperparameters evaluated.

Random Forest [Breiman 2001] was selected as the primary tree-based baseline for tabular traffic, while Logistic Regression [Friedman et al. 2010] provides a linear reference. LightGBM [Ke et al. 2017] supports the configuration-sensitivity analysis through a comparison between default and tuned settings, and the 1D-CNN [Goodfellow et al. 2016] represents deep-learning architectures commonly used in IoMT intrusion detection. For each classification scenario, the tuned LightGBM variant was selected according to validation macro-F1. The autoencoder [Goodfellow et al. 2016] provides benign-only anomaly screening and was retrained using five seeds (40–44), with the benign train/validation split fixed at seed 42. At inference, a flow is considered anomalous when its reconstruction error exceeds τ, defined as the 85th percentile (P_{85}) of the benign validation reconstruction-error distribution.

3.5. Evaluation Protocol

Evaluation combines detection, statistical, operational, and explainability evidence. Supervised detection uses accuracy, macro-F1, Cohen's κ, and per-class/phase precision

and recall. Macro-F1 assigns equal weight to all classes, while the accuracy–κ gap helps reveal whether high accuracy mainly reflects dominant classes. The binary autoencoder is reported with attack recall, precision, benign FPR and #FP, phase-level recall, AUROC, and AP for both attack and benign classes. All autoencoder metrics are summarized as mean $\pm$ sample SD across five retrainings.

Fixed-run supervised predictions receive 95% bootstrap CIs (2,000 iterations, seed 42) to assess stability under test-set resampling, not training-run variability. Pairwise McNemar tests with continuity correction [McNemar 1947] are applied only among supervised classifiers that share the same label space. The autoencoder is excluded from those comparisons and instead receives five-run SD and 95% Student-t intervals, which estimate retraining stability under the fixed benign split.

Operational measurements use pre-trained models on 45-feature flows. Single-flow latency averages 1,000 batch-one predictions after 100 warm-ups; throughput is $N/\bar{t}$ over five complete passes of the $N = 1{,}614{,}182$ test flows in batches of 512. Timings exclude packet capture, flow reconstruction, feature extraction, model loading, and queueing. Classical models run on CPU; neural models are timed on CPU and GPU; the autoencoder path includes reconstruction, per-flow MSE, and the P_{85} decision. Serialized size and training time describe model footprint and construction cost but are not end-to-end deployment measurements. SHAP and LIME support attribution inspection only, not causal explanations or deployment claims.

3.6. Experimental Reproducibility

All experiments run on AMD Ryzen 9 7940HS, 48 GB RAM, NVIDIA RTX 4060 (8 GB), Ubuntu, Python 3.12, using TensorFlow 2.x, scikit-learn, and LightGBM. Seed 42 for supervised models, stratified 80/20 split, StandardScaler on training data only. Derived results and supplementary evaluation artifacts will be made available upon publication to support independent inspection of the reported findings.

4. Results

The research questions introduced in Section 1 are addressed as RQ1 (§4.1), RQ2 (§4.2), and RQ3 (§4.3); Section 4.4 adds operational context.

4.1. Detection Capability by Attack Phase (RQ1)

Table 4 summarizes detection performance across three classification scenarios with increasing granularity. The binary setting indicates high overall performance for all models, whereas the 19-class setting exposes larger differences in fine-grained attack discrimination, which is relevant for incident triage.

RF achieves the highest F1-macro in the multiclass settings, with 0.95 in the 6-class scenario and 0.9207 in the 19-class scenario. Its small accuracy and κ gap in the 19-class setting suggest limited sensitivity to majority-class dominance. In contrast, the 1D-CNN reaches 0.9783 accuracy but 0.7488 F1-macro, indicating that accuracy alone can hide degradation on less frequent classes. As label granularity increases from 2-class to 19-class, RF loses 6.4 percentage points of F1-macro, compared with 20.9 percentage points for the 1D-CNN and 22.6 points for LightGBM Tuned. This result is consistent with prior evidence on the strength of tree-based ensembles for tabular data [Grinsztajn et al. 2022].

Table 4. Classification results across scenarios. Bold indicates the best result per scenario.

Scenario	Model	Accuracy	F1-macro	κ	Acc–κ
2-class	Random Forest	**0.9986**	0.9847	0.9693	0.029
	LightGBM Tuned	**0.9986**	**0.9853**	**0.9705**	0.028
	1D-CNN	0.9960	0.9576	0.9153	0.081
	Logistic Regression	0.9954	0.9498	0.8996	0.096
6-class	Random Forest	**0.9986**	**0.9500**	**0.9971**	0.002
	LightGBM Tuned	0.9981	0.9348	0.9961	0.002
	1D-CNN	0.9918	0.8506	0.9834	0.008
	Logistic Regression	0.7691	0.7111	0.4382	0.331
19-class	Random Forest	**0.9967**	**0.9207**	**0.9961**	**0.001**
	LightGBM Tuned	0.8281	0.7598	0.8007	0.027
	1D-CNN	0.9783	0.7488	0.9748	0.004
	Logistic Regression	0.7604	0.5335	0.7129	0.047

4.1.1. Disaggregation by Kill Chain Phase

Table 5 reports Precision, Recall, and F1 by kill-chain category, using the CICIoMT2024 taxonomy to group the 18 attack categories plus benign traffic. Each value is the unweighted mean of the corresponding per-class metric within the category; detailed per-class results and supports are reported in Table 6.

Table 5. Phase-level detection in the 19-class scenario; values are unweighted per-class means from Table 6. Best F1 in bold.

Kill-chain objective	Category	Metric	RF	1D-CNN	LGBM-T	LR
Reconnaissance	Recon n=4 sup. 27.7k	Precision	0.9014	0.3974	0.7476	0.5554
		Recall	0.6556	0.5554	0.6779	0.3736
		F1	**0.7362**	0.3979	0.7012	0.3957
Access & exploitation	Spoofing n=1 sup. 1.7k	Precision	0.7189	0.1546	0.6005	0.3555
		Recall	0.8343	0.7890	0.5361	0.4329
		F1	**0.7723**	0.2585	0.5665	0.3904
Impact	DoS n=4 sup. 417k	Precision	0.9997	0.9909	0.8515	0.5674
		Recall	0.9996	0.9952	0.7560	0.1626
		F1	**0.9996**	0.9930	0.6606	0.1897
	DDoS n=4 sup. 1.07M	Precision	0.9995	0.9973	0.9316	0.7530
		Recall	0.9992	0.9947	0.9093	0.9810
		F1	**0.9993**	0.9960	0.9051	0.8503
	MQTT n=5 sup. 63.7k	Precision	0.9804	0.7927	0.9059	0.8417
		Recall	0.9467	0.7946	0.7870	0.6379
		F1	**0.9612**	0.7086	0.7669	0.6223

Three findings emerge with direct implications for hospital defense:

- **Early-stage detection gap**. RF achieves an average recall of 0.656, compared with 0.555 for the CNN, a substantial difference at the kill chain phase, where detection enables prevention rather than merely incident response. Tuned LightGBM also

achieves competitive recon recall (0.6779), slightly above RF (0.6556), while the CNN lags substantially. The gap widens for ARP Spoofing (F1 = 0.772 vs. 0.259), indicating substantially lower coverage of spoofing-related traffic by the CNN configuration in a category associated with early-stage adversarial positioning. Note that in the dataset, the *Access & Exploitation* phase is represented exclusively by ARP Spoofing (1,744 test samples, 0.1% of the test set); conclusions for this phase therefore refer specifically to ARP Spoofing and should not be interpreted as comprehensive coverage of the broader exploitation phase. Similarly, Recon-Ping Sweep has only 186 test samples; per-class conclusions for that attack carry additional uncertainty.

- **Volumetric attacks.** Models achieve higher DDoS precision, confirming that these high-volume attacks are detectable regardless of model complexity. This is precisely why global accuracy dominated by DDoS flows provides an incomplete picture of defensive capability.
- **Practical implication.** Overall, the phase-level analysis shows that strong aggregate accuracy is largely aligned with performance on high-volume impact classes and does not necessarily imply reliable coverage of lower-support early-stage categories. A model selected only by global accuracy may therefore appear suitable while leaving reconnaissance or spoofing weakly covered. This motivates the per-class analysis in Table 6 and the deployment guidance in Section 7.

Table 6. Per-Class Detection Metrics Precision (P), Recall (R), and F1 for the four main supervised models

Class	Support	RF			1D-CNN			LGBM-T			LR		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Benign	37,607	0.9699	0.9819	0.9758	0.9877	0.7922	0.8792	0.9658	0.9697	0.9677	0.8781	0.9060	0.8919
DDoS-ICMP	349,699	0.9995	0.9997	0.9996	0.9982	0.9982	0.9982	0.7273	0.9999	0.8421	0.7815	0.9982	0.8766
DDoS-SYN	172,397	0.9999	0.9979	0.9989	0.9972	0.9884	0.9928	0.9999	0.9992	0.9996	0.8157	0.9329	0.8703
DDoS-TCP	182,598	0.9998	0.9998	0.9998	0.9947	0.9974	0.9960	0.9999	1.0000	1.0000	0.6904	0.9987	0.8164
DDoS-UDP	362,070	0.9987	0.9995	0.9991	0.9990	0.9947	0.9968	0.9992	0.6379	0.7787	0.7244	0.9942	0.8381
DoS-ICMP	98,432	0.9997	0.9991	0.9994	0.9925	0.9924	0.9925	0.4231	0.9995	0.5945	0.5332	0.0060	0.0120
DoS-SYN	98,595	0.9998	0.9998	0.9998	0.9900	0.9944	0.9922	0.9999	0.9999	0.9999	0.8575	0.6362	0.7304
DoS-TCP	82,096	0.9999	0.9999	0.9999	0.9941	0.9953	0.9947	1.0000	0.9999	1.0000	0.7477	0.0040	0.0080
DoS-UDP	137,553	0.9995	0.9994	0.9994	0.9868	0.9986	0.9927	0.9831	0.0246	0.0480	0.1312	0.0044	0.0085
MQTT-DDoS-Connect_Flood	41,916	0.9999	1.0000	0.9999	0.9968	0.9960	0.9964	0.9998	1.0000	0.9999	0.9863	0.9981	0.9921
MQTT-DDoS-Publish_Flood	8,416	0.9999	0.8952	0.9446	0.9448	0.1199	0.2128	1.0000	0.1456	0.2541	0.9871	0.1185	0.2115
MQTT-DoS-Connect_Flood	3,131	1.0000	0.9987	0.9994	0.9270	0.9850	0.9551	1.0000	0.9987	0.9994	0.9698	0.9029	0.9352
MQTT-DoS-Publish_Flood	8,505	0.9067	1.0000	0.9511	0.5315	0.9995	0.6940	0.5421	1.0000	0.7030	0.5291	0.9960	0.6911
MQTT-Malformed_Data	1,747	0.9953	0.8397	0.9109	0.5632	0.8724	0.6845	0.9878	0.7905	0.8782	0.7361	0.1740	0.2815
Recon-OS_Scan	3,834	0.8640	0.6364	0.7330	0.1773	0.4857	0.2598	0.9422	0.6161	0.7450	0.6126	0.0610	0.1110
Recon-Ping_Sweep	186	0.9197	0.6774	0.7802	0.2131	0.8925	0.3440	0.6699	0.7527	0.7089	0.7222	0.4892	0.5833
Recon-Port_Scan	22,622	0.9372	0.9893	0.9625	0.8919	0.5891	0.7095	0.9123	0.9908	0.9499	0.8286	0.9404	0.8810
Recon-VulScan	1,034	0.8847	0.3191	0.4691	0.3072	0.2544	0.2783	0.4661	0.3520	0.4011	0.0580	0.0039	0.0073
Spoofing	1,744	0.7189	0.8343	0.7723	0.1546	0.7890	0.2585	0.6005	0.5361	0.5665	0.3555	0.4329	0.3904

4.2. Configuration Sensitivity and Detection Degradation (RQ2)

This section examines whether configuration alone, independent of architecture, can create equivalent blind spots. LightGBM was selected for this analysis because it allows a controlled internal comparison between a default configuration and tuned variants while keeping the same model family, isolating configuration effects from architectural differences. Table 7 quantifies the detection impact of deploying LightGBM with default parameters and tuned settings under the same feature representation, preprocessing pipeline, and test protocol.

Table 7. LightGBM F1-macro under default and tuned configurations. ΔF1-m is the absolute gain, and the Acc/F1 ratio reflects divergence between aggregate accuracy and macro-averaged performance.

Scenario	F1-m (Default)	F1-m (Tuned)	Δ F1-m	Acc/F1 Ratio
2-class	0.9826	**0.9853**	+0.003	1.02
6-class	0.8805	**0.9348**	+0.054	1.13
19-class	0.2256	**0.7598**	+0.534	2.37

Configuration has little effect on binary detection but becomes increasingly consequential at finer classification granularities. Tuning improves F1-macro by 0.003, 0.054, and 0.534 in the 2, 6, and 19-class scenarios, respectively. Although the default model achieves 99.82% binary accuracy, its 19-class F1-macro falls to 0.2256. The Acc/F1 ratio of 2.37 indicates that aggregate accuracy may conceal uneven class-level performance, reducing the specificity available for incident triage.

This improvement cannot be assigned to a single hyperparameter, since tuning jointly changes the number of estimators, learning rate, number of leaves, early stopping, and imbalance handling. The experiment therefore demonstrates sensitivity to the overall configuration, while isolating the contribution of `is_unbalance=True` or another parameter would require a controlled ablation. These results support configuration verification and the use of macro-averaged and per-class metrics as part of IDS deployment validation, particularly in highly imbalanced multiclass settings.

4.3. Benign-Only Anomaly Detection (RQ3)

The previous findings assume labeled attack data is available for training. However, hospitals face novel threats, new MQTT exploits, ransomware variants, and targeted attacks against specific device models for which no training labels are available. Table 8 presents the autoencoder's detection capability when trained exclusively on benign traffic, reported as the mean $\pm$ standard deviation across five training seeds (40–44) with the benign train/validation split held fixed (seed 42).

Table 8. Benign-only autoencoder as a binary anomaly detector over the original 19-class test partition (mean $\pm$ SD over five seeds; fixed benign train/validation split). Threshold τ is the given percentile of benign validation reconstruction errors. False positives are counted over 37,607 benign test flows.

Thr.	τ	Recall	Precision	FPR	#FP	F1-macro
P85	0.931 ± 0.076	0.994 ± 0.006	0.994 ± 0.000	0.244 ± 0.003	9181 ± 93	0.876 ± 0.044
P90	1.594 ± 0.216	0.825 ± 0.257	0.994 ± 0.003	0.173 ± 0.006	6521 ± 214	0.677 ± 0.256
P95	2.846 ± 0.386	0.277 ± 0.383	0.982 ± 0.011	0.091 ± 0.006	3442 ± 210	0.242 ± 0.279
P99	6.997 ± 0.794	0.018 ± 0.012	0.958 ± 0.030	0.023 ± 0.004	874 ± 146	0.041 ± 0.012

Threshold-independent performance reached AUC-ROC $= 0.878 \pm 0.036$, AP(attack) $= 0.991 \pm 0.002$, and AP(benign) $= 0.790 \pm 0.041$. Since attacks comprise approximately 97.7% of the test partition, AP(attack) benefits from a high prevalence baseline; AP(benign), computed using negative reconstruction error, provides a more informative view of separability. At P85, the detector achieves recall of 0.994 ± 0.006, but

Table 9. Autoencoder phase-level attack recall at $\tau =$ P85 (mean $\pm$ SD over five seeds). Macro recall is the unweighted mean of per-class recalls in each phase.

Phase	Classes	Support	Macro recall
Reconnaissance	4	27,676	0.786 ± 0.010
ARP Spoofing	1	1,744	0.729 ± 0.011
Impact: DoS	4	416,676	0.991 ± 0.016
Impact: DDoS	4	1,066,764	0.999 ± 0.002
Impact: MQTT	5	63,715	0.867 ± 0.052

its FPR of 0.244 ± 0.003 corresponds to approximately 9,181 false positives. Increasing the threshold reduces false alarms but sharply decreases recall and increases variability: recall falls to 0.277 ± 0.383 at P95 and 0.018 ± 0.012 at P99. P85 is therefore the highest-recall and most stable tested setting, although its false-positive burden indicates that additional filtering would be required for direct alert generation.

Phase-disaggregated recall (Table 9) is not uniform: the autoencoder flagged the majority of volumetric Impact attacks (DDoS 0.999 ± 0.002, DoS 0.991 ± 0.016), nonetheless, it is weaker on early-stage and low-volume categories (Reconnaissance 0.786 ± 0.010, ARP Spoofing 0.729 ± 0.011, MQTT 0.867 ± 0.052). Thus, the autoencoder provides complementary anomaly-screening evidence, but its utility depends on threshold calibration and attack characteristics. Finally, this experiment evaluates binary anomaly detection, not attack classification or zero-day detection, since all tested attacks belong to CICIoMT2024.

4.4. Operational Viability for Gateway Deployment

Operational deployment also depends on latency, throughput, model size, training cost, and availability. Table 10 reports a consolidated benchmark using the same environment, procedure, and 19-class test partition. Latency is the mean single-flow inference time, whereas throughput is measured over the full test set. Classical models run on CPU; neural models are evaluated on CPU and GPU.

Table 10. Consolidated operational profile on the 19-class test partition.

	CPU		GPU			
Model	**Lat. (ms)**	**Thr. (s^{-1})**	**Lat. (ms)**	**Thr. (s^{-1})**	**Disk (MB)**	**Train (s)**
Random Forest	15.67	417,219	–	–	783.56	192.5
Logistic Regression	0.032	4,565,432	–	–	0.01	1,154.0
LightGBM (Default)	2.57	206,904	–	–	3.78	270.4
LightGBM (Tuned)	5.22	4,888	–	–	134.21	8,087.6
1D-CNN	37.85	132,907	50.93	268,538	1.62	6,293.4
Autoencoder	37.74	496,654	51.75	217,241	0.22	19.77 ± 5.14

The classical models present distinct trade-offs. Logistic Regression is the fastest and smallest, but its lower 19-class macro-F1 limits fine-grained detection. Default LightGBM is compact and low-latency, although its weak multiclass performance shows that efficiency alone is insufficient. Tuning improves detection but increases model size

to 134.21 MB and training time to 8,087.6 s, while reducing throughput to 4,888 flows/s. Random Forest provides the strongest supervised performance with 417,219 flows/s, but its 783.56 MB artifact is the largest.

GPU execution increases batched throughput for the CNNs but not single-flow latency, for which CPU execution is faster under the evaluated procedure. The autoencoder also achieves higher throughput on CPU than GPU, indicating no inference advantage from acceleration for this architecture. Its 0.22 MB model trains in 19.77 ± 5.14 s and reaches recall of 0.994 ± 0.006 at P_{85}, but the corresponding FPR of 0.244 ± 0.003 requires threshold calibration or downstream filtering.

Overall, Random Forest offers the strongest supervised detection-throughput balance, with footprint as its main constraint. The autoencoder is lightweight but produces a substantial false-positive burden, while tuned LightGBM incurs the highest computational cost. These workstation-level measurements exclude feature extraction, memory usage, energy, concurrency, and queueing; validation on representative gateway hardware remains necessary.

4.5. Contextualization against prior CICIoMT2024 studies.

Table 11 compares reported headline metrics with the present study. The comparison is descriptive and does not constitute a ranking, as train/test splits, preprocessing, sampling strategies, and F1 averaging conventions differ. Its purpose is to contextualize reporting practices and motivate phase-aware analysis.

Table 11. Headline CICIoMT2024 detection metrics from prior IDS studies.

Study	Model	2-class		6-class		19-class	
		Acc	F1	Acc	F1	Acc	F1
Dadkhah et al.	Random Forest	0.996	0.961	0.735	0.676	0.733	0.551
	DNN	0.996	0.952	0.734	0.665	0.729	0.522
Sohail et al.	XGBoost	–	–	–	–	0.950	–
Mohammadi et al.	1D-CNN	0.99	0.99	0.99	0.99	0.99	0.98
Kavkas & Yildiz	DNN	0.99	0.99	0.78	0.75	0.77	0.71
	LSTM	0.99	0.99	0.79	0.76	0.78	0.71
Akar et al.	L2D2 (LSTM)	–	–	–	–	0.98	0.98
Yacoubi et al.	RF (XAI)	0.999	0.999	–	–	–	–
Chandekar et al.	Stacking ens.	0.96	–	–	–	–	–
This work	Random Forest	0.9986	0.9847	0.9986	0.9500	**0.9967**	**0.9207**
	LightGBM (Tuned)	0.9986	0.9853	0.9981	0.9348	0.8281	0.7598
	LightGBM (Default)	0.9984	0.9826	0.9906	0.8805	0.5352	0.2256
	1D-CNN	0.9960	0.9576	0.9918	0.8506	0.9783	0.7488
	Autoencoder†	0.9882	0.8756	–	–	–	–

Notes: "–" denotes scenarios not reported by the cited study. †The autoencoder is trained on benign-only traffic and operates as an anomaly screen; the 2-class entries are five-seed mean accuracy and macro-F1 for benign-versus-attack discrimination at the P_{85} reconstruction-error threshold (§4.3).

5. Statistical Validation and Feature Importance

Deploying an IDS in a hospital requires confidence that observed differences are genuine rather than artifacts of test-set composition, and that model decisions are grounded in

legitimate network behavior. This section addresses both concerns.

Statistical Confidence. Table 12 presents 95% bootstrap confidence intervals (2000 iterations, resampling predictions without retraining) for the 19-class scenario. The intervals are narrow (CI width ≤ 0.007 for F1-macro), confirming that results are stable across test-set compositions. Critically, the RF confidence interval for F1-macro [0.9172, 0.9240] does not overlap with those of LightGBM Tuned [0.7561, 0.7633] or the CNN [0.7459, 0.7517], indicating consistently higher multiclass performance for RF under the evaluated conditions.

Table 12. Bootstrap 95% CI for the supervised classifiers in the 19-class scenario. The benign-only autoencoder is excluded because it is characterized by its across-seed variability (Table 8) rather than by prediction-level bootstrap.

Model	Accuracy [95% CI]	F1-macro [95% CI]	κ [95% CI]
Random Forest	0.9967 [.9966, .9967]	0.9207 [.9172, .9240]	0.9961 [.9960, .9962]
LightGBM Tuned	0.8281 [.8276, .8287]	0.7598 [.7561, .7633]	0.8007 [.8001, .8014]
1D-CNN	0.9783 [.9781, .9786]	0.7488 [.7459, .7517]	0.9748 [.9746, .9751]
Logistic Reg.	0.7603 [.7597, .7610]	0.5335 [.5294, .5371]	0.7129 [.7122, .7136]

McNemar's test with continuity correction, computed pairwise among the supervised classifiers within each scenario (all four models share the same label space, and the benign-only autoencoder is not included), yields $p < 0.001$ for all supervised model pairs, with one exception: RF vs. LightGBM Tuned in the binary scenario ($\chi^2 = 0.076$, $p = 0.783$). Because McNemar's statistic is sensitive to sample size and can flag negligible differences as significant on a test set of this magnitude, the practical effect size is read from the separation of the F1-macro bootstrap intervals in Table 12: RF [0.9172, 0.9240] does not overlap LightGBM Tuned [0.7561, 0.7633] or the CNN [0.7459, 0.7517], a gap of roughly 16–17 percentage points. No statistically significant difference was detected between RF and tuned LightGBM in the binary scenario, while RF is both statistically and materially superior when fine-grained classification is required.

Cohen's Kappa further reveals that RF's high accuracy is not inflated by majority-class dominance: its accuracy–κ gap is only 0.0005 in the 19-class scenario, indicating consistency between aggregate accuracy and class-balanced agreement.

Autoencoder Training Stability. The statistical evidence for the autoencoder is of a different nature and is reported separately to avoid conflation with the supervised intervals above. The bootstrap intervals in Table 12 quantify *test-set* resampling stability of a single fixed run, whereas the autoencoder is characterized by its *training-run* variability across the five seeds (40–44), summarized by 95% Student-t confidence intervals over those five retrainings (benign train/validation split held fixed at seed 42). Under this protocol, the threshold-independent AUC-ROC is 0.878 (95% CI $[0.833, 0.923]$) and the P_{85} F1-macro is 0.876 (95% CI $[0.822, 0.930]$); the corresponding P_{85} attack recall is 0.994 (95% CI $[0.986, 1.000]$, upper bound truncated at the metric ceiling). These intervals are wider than the supervised bootstrap intervals, as expected from the small number of retrainings, and should be read as an estimate of training-run stability rather than as a large-sample significance test.

Feature-Attribution Sanity Check. SHAP and LIME are used to inspect whether

the supervised models rely on features with plausible network-security meaning. This analysis is limited to feature attribution and does not constitute a causal explanation or a complete audit of model behavior. SHAP TreeExplainer was applied in the 19-class scenario using 1,000 background samples and 500 test samples. The most influential features across models are mainly related to traffic timing, connection state, and protocol indicators. No obvious identifier, ordering, or absolute timestamp feature appears among the main contributors. Detailed SHAP plots are reported in Figure 2. LIME provides a

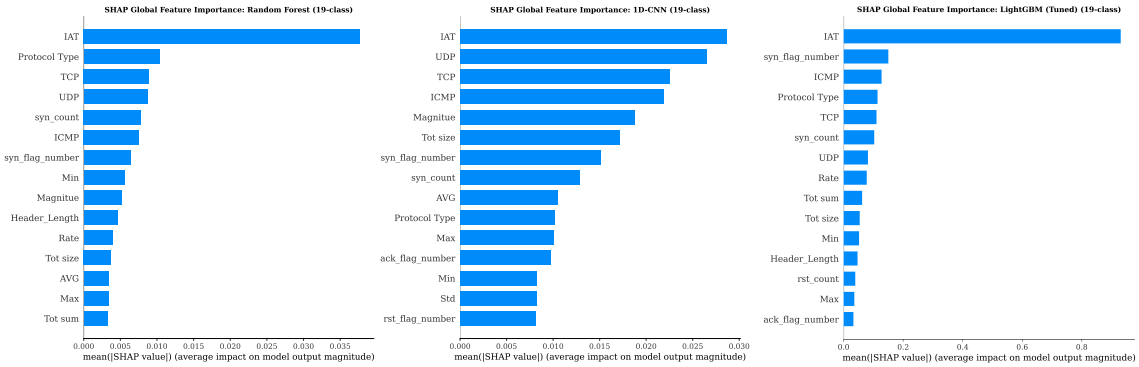

Figure 2. Global SHAP feature-attribution summaries for the 19-class scenario. From left to right: Random Forest, 1D-CNN, and tuned LightGBM.

local inspection of two Random Forest predictions for Recon-Ping_Sweep: one correctly classified instance and one misclassified instance. With 5,000 perturbations and 15 features, the explanations indicate that ARP and flow-activity indicators contribute to the correct prediction, while the misclassified instance shows weaker ARP contribution and conflicting packet-activity features. These plots are illustrative instance-level evidence and are also reported in Figure 3.

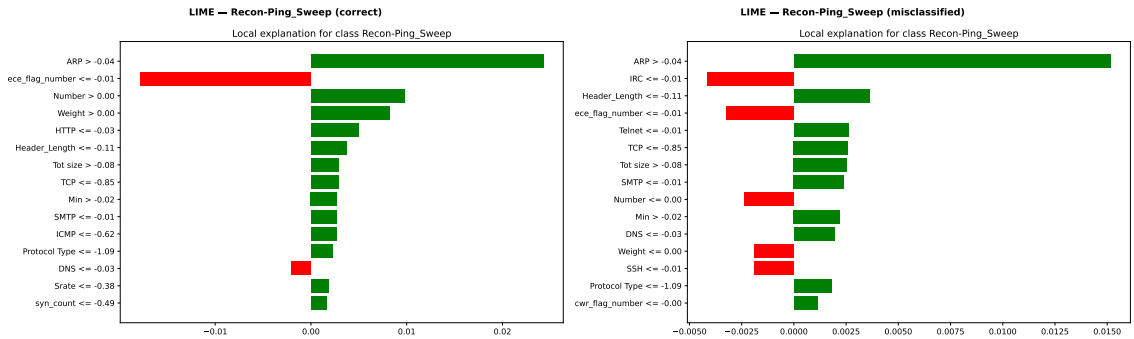

Figure 3. LIME local for two Recon-Ping_Sweep instances classified by RF.

6. Threats to Validity

Dataset scope. Experiments use only CICIoMT2024 [Dadkhah et al. 2024] and its Wi-Fi/MQTT feature-level data; BLE is excluded. Cross-dataset degradation has been reported [Doménech et al. 2025], so generalization to other hospitals, devices, and capture conditions remains unverified. **Class imbalance.** Minority support reaches 0.012% of the test set, making aggregate accuracy less representative of low-support attacks. Macro-F1, κ, per-class, and phase metrics mitigate but do not remove this limitation. **Protocol and**

operational scope. BLE requires Bluetooth-specific radio/link-layer observation, whereas the study uses Wi-Fi/MQTT flow features. Experimental setup timing covers inference on pre-extracted features and excludes capture, flow reconstruction, feature extraction, model loading, queueing, memory, energy, and concurrent traffic. **Run variability.** Autoencoder SD and Student-t CIs use five seeds. Supervised models use one seed; their bootstrap CIs quantify test-set resampling stability, not retraining variance. **Training asymmetry.** Supervised models use the full labeled partition, whereas the autoencoder uses $\sim$165k benign flows, preventing performance differences from being attributed solely to model capability. **Prior-work comparison.** Table 11 is descriptive, as protocols, sampling, and F1 averaging differ; gains cannot be attributed without exact reproduction.

7. Conclusion and Deployment Implications

This paper presented a phase-aware evaluation of baseline learning models for intrusion detection in IoMT networks using CICIoMT2024. The analysis compared standard supervised models and a benign-only autoencoder across binary, coarse-grained, and fine-grained detection settings, with emphasis on macro-averaged, per-class, and attack-type evidence.

Under the evaluated configurations, Random Forest achieved higher recall on early-stage attack categories than the 1D-CNN, with a 10.1 pp recall advantage on reconnaissance (0.656 vs. 0.555) and a 51.3 pp F1 advantage on spoofing (0.772 vs. 0.259). It also showed a smaller macro-F1 decrease when moving from binary to 19-class detection (6.4 pp) than the CNN (20.9 pp). These results are consistent with prior evidence that tree-based models remain strong baselines for tabular data [Grinsztajn et al. 2022], while the contribution here is to show how these differences manifest under attack-type and phase-aware IoMT evaluation.

The LightGBM results show that configuration can substantially affect fine-grained detection. The accuracy/F1-macro ratio increases from 1.02 in the binary setting to 2.37 in the 19-class setting, while tuned LightGBM improves macro-F1 by 53.4 pp over the default configuration in the 19-class task. This does not attribute the gain to a single parameter; rather, it indicates that gradient-boosting IDSs should be assessed under the intended class granularity using macro-averaged and per-class metrics. The benign-only autoencoder provides complementary anomaly-screening evidence. At P85, it achieved attack recall 0.994 ± 0.006 and AUROC 0.878 ± 0.036 across five seeds, but incurred a benign false-positive rate of 0.244 ± 0.003, without using attack samples during training. These results should not be interpreted as attack-type classification or strict zero-day validation, since all test attacks belong to CICIoMT2024 categories. Instead, they support reconstruction error as a complementary screening signal, subject to threshold-specific false-positive burden and target-hardware validation.

Three implications follow for IoMT IDS evaluation and deployment planning. First, aggregate accuracy should be complemented with macro-F1 and class- and phase-disaggregated metrics, particularly under severe imbalance. Second, model configurations should be validated at the classification granularity intended for deployment, since strong binary performance may not transfer to fine-grained incident triage. Third, anomaly screening and supervised classification should be assessed together with end-to-end operational costs on representative gateway hardware, including packet capture, flow reconstruction, feature extraction, memory, energy, concurrency, and queueing.

Two directions extend this work toward production readiness: evaluation with real hospital traffic to measure false positive rates under operational conditions, and investigation of RF+Autoencoder ensemble voting where anomaly confidence modulates classification thresholds.

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

Adji, L. H. E., Cunha, L. F. I., and Machado, R. C. S. (2025). Detecção de botnets em dispositivos IoT utilizando análise de consultas DNS com one-class SVM. In *Anais do SBSeg 2025: Artigos Completos*, pages 1–17. Sociedade Brasileira de Computação.

Akar, G., Sahmoud, S., Onat, M., Cavusoglu, Ü., and Malondo, E. (2025). L2d2: A novel lstm model for multi-class intrusion detection systems in the era of iomt. *IEEE Access*, 13:7002–7013.

Ayad, A. G., Sakr, N. A., and Hikal, N. A. (2025). Fog-empowered anomaly detection in iot networks using one-class asymmetric stacked autoencoder. *Cluster Computing*, 28(8).

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chandekar, P. B., Mehta, M. S., and Chandan, S. (2025). Enhanced anomaly detection in iomt networks using ensemble ai models on the ciciomt2024 dataset. arXiv preprint arXiv:2502.11854.

Dadkhah, S., Neto, E. C. P., 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*, 28:101351.

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 SBSeg 2025: Artigos Completos*, pages 1–17. Sociedade Brasileira de Computação.

Doménech, J., León, O., Siddiqui, M. S., and Pegueroles, J. (2025). Evaluating and enhancing intrusion detection systems in iomt: the importance of domain-specific datasets. *Internet of Things*, 32:101631.

Ehrenfeld, J. M. (2017). WannaCry, cybersecurity and health information technology: A time to act. *Journal of Medical Systems*, 41(7):104.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, MA, USA. Chapter 14: Autoencoders.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Hutchins, E. M., Cloppert, M. J., and Amin, R. M. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. In *Proceedings of the 6th International Conference on Information Warfare and Security (ICIW 2011)*, pages 113–125, Washington, DC, USA. Academic Conferences and Publishing International Ltd. Conference held 17–18 March 2011 at George Washington University; also issued as Lockheed Martin white paper.

Kavkas, N. C. and Yildiz, K. (2025). Enhancing iomt security with deep learning based approach for medical iot threat detection. In *IEEE International Symposium on Digital Forensic and Security*, ISDFS, pages 1–6.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 3146–3154, Long Beach, CA, USA. Curran Associates, Inc.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157. PMID: 20254758.

MITRE ATT&CK (2025). Enterprise tactics: Reconnaissance and impact. `https://attack.mitre.org/tactics/`. Accessed: 2026-05-11.

Mohammadi, A., Ghahramani, H., Asghari, S. A., and Aminian, M. (2024). Securing healthcare with deep learning: A cnn-based model for medical iot threat detection. In *19th Iranian Conference on Intelligent Systems*, ICIS, pages 168–173.

Ruellan, E., Paquet-Clouston, M., and Garcia, S. (2024). Conti inc.: Understanding the internal discussions of a large ransomware-as-a-service operator with machine learning. *Crime Science*, 13(16).

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. In *International Conference on Digital Transformation*, ICoDT2, pages 1–6. IEEE.

Uddin, M. A., Chu, N. H., and Rafeh, R. (2025). A hierarchical ids for zero-day attack detection in internet of medical things networks. arXiv preprint arXiv:2508.10346.

Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., and Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access*, 7:41525–41550.

Yacoubi, M., Moussaoui, O., and Drocourt, C. (2026). Enhancing iomt security with explainable machine learning: A case study on the ciciomt2024 dataset. In *Connected Objects, Artificial Intelligence, Telecommunications and Electronics Engineering*, pages 249–254, Cham. Springer Nature Switzerland.