Journal of Intelligent Systems and Internet of Things

ISSN: 2690-6791 Vol: 19, No. 01, 2027 PP. 35–46

Article DOI: https://doi.org/10.54216/JISIoT.190103



FULL LENGTH ARTICLE

An Information-Theoretic Framework for Evaluating Class Imbalance and Rare-Attack Observability in IoT Security Benchmarks

Firas Zobary[1,*]

[1] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, Hubei, China

Email: firaszobary@whut.edu.cn

Received: August 02, 2026 Revised: September 16, 2026 Accepted: November 04, 2026 ⋆ Corresponding author

ABSTRACT

IoT intrusion-detection studies often treat a benchmark as adequate once it is large, recent, and widely used. That assumption can fail. This study audits five public IoT and IIoT security benchmarks—CICIoT2023, Edge-IIoTset, ToN-IoT, BoT-IoT, and IoT-23—from their published class populations and introduces no synthetic classifier results; the audit combines Shannon entropy, Hill diversity, Simpson concentration, Gini inequality, rank-frequency tail fitting, rare-class observation probability, Wilson uncertainty, and a constrained power-tempered sampling model. Across the five corpora, record counts range from 461,043 to 73,370,443, normalized entropy from 0.421 to 0.600, and effective Shannon diversity from 2.20 to 3.98 classes. Size is a poor proxy for label adequacy. IoT-23 is the clearest case: its 6.05 million records include a singleton class, so a random 1% sample has only about a 1% probability of observing that label, while the worst-case Wilson half-width for a class-conditioned proportion is 0.445. We then define $q_i(\tau) = p_i^\tau / \sum_k p_k^\tau$ and prove that its entropy is non-increasing in τ; at a 1% training budget with a target of 25 expected observations in the rarest class, the minimum-flattening exponent is 1.000 for CICIoT2023, 0.855 for Edge-IIoTset, 0.807 for ToN-IoT, 0.954 for BoT-IoT, and 0.450 for IoT-23. The last case needs an approximately 2,500-fold uplift of the singleton label. Such an uplift changes training exposure to one observation; it cannot create independent evidence. The resulting audit protocol therefore separates training rebalancing from evaluation validity and gives quantitative checks that can be applied before IoT benchmark results are interpreted as class-wide security evidence.

Keywords: Internet of Things ▪ Intrusion detection ▪ Information theory ▪ Class imbalance ▪ Benchmark adequacy ▪ Rare attacks ▪ Entropy ▪ Tempered sampling

1. INTRODUCTION

IoT security research has accumulated benchmarks faster than it has accumulated rules for deciding what those benchmarks can support. Large corpora make model fitting convenient, allow deep architectures to consume millions of records, and give headline accuracy values an appearance of numerical stability, yet the label distribution can tell a different story: a corpus with seventy million flows may still allocate only a few thousand observations to a rare class, while another corpus one hundred times smaller can provide enough independent evidence for materially tighter class-conditioned estimates. Size is one coordinate of adequacy. Rare-class support is another.

The distinction matters for intelligent IoT security systems because recent work has moved intrusion detection toward

© 2027 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

boosting, explainable models, federated learning, and distributed recurrent architectures [1, 2, 3, 4, 5]. Those studies mainly ask an algorithmic question: given a benchmark and a split, which model detects the labeled traffic most effectively? The benchmark is usually treated as fixed. Multi-dataset studies can make the same assumption even when CICIoT2023, BoT-IoT, and Edge-IIoTset are compared side by side [6], although a thin label tail can make a random sample omit an attack altogether and can turn one or two independent observations into thousands of repeated training presentations after aggressive oversampling. The substrate therefore needs its own test.

Imbalanced learning is not a new problem. He and Garcia [7] described the general statistical and algorithmic difficulties of skewed class populations, Saito and Rehmsmeier [8] showed why common evaluation views can be misleading under imbalance, and resampling or loss correction has been studied from SMOTE [9] through effective-number weighting [10] and logit adjustment [11]; those methods operate mainly after a dataset has been accepted for learning. This paper asks an earlier question. What does the empirical label geometry permit the researcher to infer before a classifier is fitted?

The answer is developed with five datasets that span current IoT security practice. CICIoT2023 contributes 46.69 million records across eight coarse categories [12]. Edge-IIoTset contributes 2.22 million records across fifteen classes and has been used for centralized and federated IIoT research [13, 14]. The processed ToN-IoT network distribution supplies ten classes with a much smaller total population [15, 16]. BoT-IoT provides more than 73 million records but an extremely thin information-theft class [17]. IoT-23 is smaller than CICIoT2023 and BoT-IoT, yet its fine-grained label tail reaches one record [18, 19]; these benchmarks are kept distinct because their taxonomies operate at different semantic levels, and the comparison is restricted to the distribution of empirical support over the labels that each benchmark asks a model to distinguish.

The study makes three contributions, each tied to a different stage of benchmark use. A benchmark-adequacy calculus combines entropy, Hill numbers, concentration, class-count inequality, rank-frequency decay, and finite-sample observability; an analytic sampling family then moves continuously from the native distribution to the uniform distribution and admits a closed derivative for entropy; finally, an evaluation rule separates the two operations that are often mixed in practice, because training distributions may be altered to expose a learner to rare attacks while test evidence must still be judged from independent class counts and their uncertainty. That separation matters.

The paper omits classifier accuracy. That omission is deliberate. A model comparison would mix architecture, preprocessing, feature leakage, split design, and class geometry, making it harder to isolate the question under study; every numerical result below is computed from published class counts or from equations applied to those counts. No synthetic attacks, predicted labels, or simulated confusion matrices are introduced.

Table 1 places the five benchmark populations on the same accounting basis. The "analysis labels" column is the number of labels used in the distributional audit, which is why it may differ from finer attack taxonomies in the original repositories.

2. FROM IMBALANCE TO BENCHMARK ADEQUACY

Class imbalance is usually introduced as a learning problem: the majority class dominates the loss, the minority class contributes fewer gradient updates, and a decision boundary drifts toward the population that pays most of the empirical risk. That account is correct but incomplete. A benchmark also functions as a measurement instrument. Its class counts determine how often a rare event can enter a random subset, how tightly class-conditioned recall or precision can be estimated, and how much independent variation is available before any algorithmic correction is applied; the distinction between learning imbalance and measurement insufficiency runs through the analysis below.

Shannon's entropy [20] supplies the first coordinate because it measures uncertainty in the label identity of a randomly drawn record; Hill [21] recasts entropy as an effective number of classes, while Simpson concentration [22] gives more weight to the dominant probabilities; these quantities answer related questions without collapsing to the same number. A benchmark with $K = 15$ labels may have the Shannon diversity of only three or four equally represented classes; another with $K = 10$ may have an effective diversity close to four; raw richness and effective richness then separate sharply.

Information theory also gives a useful reference distribution. If $u_i = 1/K$, the divergence $D_{\mathrm{KL}}(p\|u) = \ln(K) - H(p)$ is exactly the entropy deficit from uniformity [23]. Nothing in this study says that a security benchmark should be uniform. Real attacks are not uniform, and a benchmark may be designed to stress a particular threat; the divergence is used as a geometric reference, with no claim that uniform class mass is a target. The same caution applies to Gini inequality and the majority-to-minority ratio: each describes the allocation of benchmark support while leaving real-world traffic prevalence as a separate empirical question.

The second coordinate is observability. Suppose a class has probability p_i in a benchmark and m records are sampled independently under the native distribution. The probability that the class is absent is $(1 - p_i)^m$. This is elementary, but it is operationally decisive when studies train on 0.1%, 1%, or 5% of a very large corpus for computational convenience; a million-row sample can sound substantial while giving almost no chance of seeing a singleton label from a six-million-record parent corpus.

Evaluation adds a third coordinate. A rare class may be present and still be too small for a precise class-conditioned estimate, which is why Wilson's score interval [24] is used here instead of the textbook Wald interval; it behaves better at small n and near the boundaries, and at the variance-maximizing estimated proportion of one half, 97 independent test records give a 95% Wilson half-width of about 0.098 while 385 reduce it to about 0.050. These counts are reference points, not adequacy thresholds. They make uncertainty visible.

Existing remedies attack another layer. SMOTE synthesizes minority samples from local neighborhoods [9], class-balanced losses change how counts enter the loss [10], and

Table 1. Benchmark provenance and analytical boundary.

Benchmark	Records	Labels	Setting	Count source
CICIoT2023	46,686,579	8	IoT traffic; 105-device testbed	Coarse category totals reported by Neto et al. [12]
Edge-IIoTset	2,219,201	15	IoT/IIoT traffic and system sources	Published 15-class processed distribution [13, 14]
ToN-IoT	461,043	10	Network subset of IoT/IIoT telemetry	Standard processed network distribution [15, 16]
BoT-IoT	73,370,443	5	Botnet-oriented IoT testbed traffic	Full category counts [17]
IoT-23	6,046,623	13	Real IoT malware and benign traffic	Aggregated label counts from public corpus analysis [18, 19]

Table 2. Mathematical notation.

Symbol	Definition
j	Benchmark index
i	Class index within benchmark j
n_{ji}	Published count of class i
N_j	Total records, $\sum_i n_{ji}$
p_{ji}	Native class probability n_{ji}/N_j
K_j	Number of analyzed labels
$H(p)$	Shannon entropy
D_q	Hill diversity of order q
m	Sampling budget
τ	Tempering exponent in $[0, 1]$
$q_i(\tau)$	Tempered training probability
r	Target expected support for the rarest class

logit adjustment changes the relation between priors and decision scores [11]. Each can improve learning. None can retroactively create independent evidence for an evaluation class that was observed once; the mathematical separation between exposure and evidence is the main reason the sampling model in Section 6 is restricted to training-set construction.

The notation used throughout the paper is collected in Table 2. It is intentionally small; every later quantity can be derived from the benchmark counts and the sampling budget.

3. DATA BOUNDARY AND REPRODUCIBLE COUNT CONSTRUCTION

The analysis uses published label populations and does not require packet payloads. This choice keeps the experiment reproducible at the level required by the questions: entropy, concentration, observation probability, and sampling weights depend on class counts alone. It also avoids silently mixing versions after preprocessing. CICIoT2023 is taken at the eight-category level reported with 46,686,579 records [12]; Edge-IIoTset uses the fifteen-class distribution reported in a 2026 study that retained 2,219,201 processed instances [14], while the provenance and feature construction remain those of the original Edge-IIoTset paper [13].

ToN-IoT needs more care because multiple subsets circulate under the same family name; this study uses the standard processed network distribution with 300,000 normal flows, eight attack classes at 20,000 each, and 1,043 MITM records, giving 461,043 observations across ten labels [15, 16]; the result should be read for that processed network benchmark only. It is not an estimate of the whole ToN-IoT collection.

BoT-IoT is analyzed at five categories. The full popula-tion has 38,532,480 DDoS records, 33,005,194 DoS records, 1,821,639 reconnaissance records, 9,543 normal records, and 1,587 theft records [17]; the normal and theft classes are therefore tiny relative to the two flooding categories; IoT-23 is analyzed with the thirteen labels reported in the pub-lic 6,046,623-record aggregation [19], whose source corpus was released by Garcia, Parmisano, and Erquiaga [18]. The thinnest labels have counts of 46, 30, 13, 8, 3, and 1.

Taxonomies remain dataset-specific. DDoS in one bench-mark is kept distinct from DDoS in another because the cap-ture conditions and taxonomies differ, and a fine malware family in IoT-23 is not equated with a coarse category in CICIoT2023; the cross-dataset calculations therefore com-pare support geometry while preserving semantic differences among the labels; that restriction matters because merging labels merely to make the benchmark names line up would manufacture comparability that the sources do not claim.

Published work continues to fit increasingly capable detectors on these corpora; Baazeem [1] applies optimized boosting in a smart-IoT IDS; Assudani et al. [2] combine imbalance handling and explainability; Sleem [3] frames industrial de-tection through federated learning; Salman [4] combines re-current units with distributed clustering; and Abdulhammed et al. [5] survey the broader IDS problem in IoT; those stud-ies motivate the audit without serving as data sources. A 2026 multi-dataset threat-intelligence study further shows that cross-benchmark experimentation is now routine [6]. The current paper asks what distributional evidence those benchmark labels carry before model choice enters the pic-ture.

For exact replication, Section 10 reports every published class count used in the audit in Table 12 and the complete budget-by-target tempering matrix in Table 13. Those two tables are sufficient to recompute the sensitivity results from the stated equations without packet payloads, trained models, or simulated classifier outputs.

4. MATHEMATICAL AUDIT OF LABEL GEOMETRY

Let $p = (p_1, \ldots, p_K)$ denote a benchmark label distribution. Shannon entropy is

$$H(p) = -\sum_{i=1}^{K} p_i \ln p_i, \tag{1}$$

and its normalized form $H^*(p) = H(p)/\ln(K)$ lies in $[0,1]$. The corresponding Hill numbers are

$$D_0 = K, \qquad D_1 = \exp(H(p)), \qquad D_2 = \frac{1}{\sum_i p_i^2}. \tag{2}$$

D_0 counts labels, D_1 is the number of equally likely labels with the same Shannon entropy, and D_2 is the inverse Simpson concentration. The Herfindahl-Hirschman concentration $\mathrm{HHI} = \sum_i p_i^2$ is therefore $1/D_2$. Against a uniform reference $u_i = 1/K$,

$$D_{\mathrm{KL}}(p\|u) = \sum_i p_i \ln\left(\frac{p_i}{u_i}\right) = \ln(K) - H(p). \tag{3}$$

The equality makes the entropy deficit directly interpretable as information divergence from equal class mass, while leaving the realism of a uniform traffic distribution as a separate empirical issue and measuring only how far the benchmark support moves away from that mathematical reference.

Count inequality is also summarized with the finite-sample Gini coefficient. For ordered counts $n_{(1)} \le \cdots \le n_{(K)}$,

$$G = \frac{2\sum_i i n_{(i)}}{K\sum_i n_{(i)}} - \frac{K+1}{K}. \tag{4}$$

The maximum-to-minimum ratio is retained because it is easy to read even though two classes determine it and a singleton can make it enormous; to describe the whole tail instead of only its endpoints, the sorted probabilities $p_{(r)}$ are fitted on a log-log scale:

$$\ln p_{(r)} = a - \alpha \ln r + \varepsilon_r, \qquad r = 1, \ldots, K. \tag{5}$$

The fitted exponent α is descriptive. With only five to fifteen labels it should not be read as proof of a power law; a large α simply records steep rank decay across the observed label support, while R^2 tells how well that one-parameter slope follows the ordered probabilities.

Table 3 reports the resulting geometry. Figure 1 plots the same distributions by rank, making the different tail shapes visible before any sampling rule is imposed.

The headline result is a decoupling between N and effective diversity: CICIoT2023 contains 46.69 million records with $D_1 = 2.40$ across eight labels, while the still larger BoT-IoT corpus contains 73.37 million records and reaches only $D_1 = 2.21$ across five categories. ToN-IoT is the smallest benchmark in Table 3 yet reaches $D_1 = 3.98$, the highest value in the comparison, so the ordering by effective diversity is plainly different from the ordering by corpus size. Size does not settle the question. Support allocation does.

Figure 2 compares D_0, D_1, and D_2 directly. The gap from D_0 to D_1 is substantial in every benchmark and especially pronounced for Edge-IIoTset, where fifteen nominal labels behave like only 3.35 equally likely classes under Shannon diversity. Figure 3 adds Gini inequality to normalized entropy. ToN-IoT occupies the least imbalanced corner of this five-dataset comparison, while IoT-23 combines $H^* = 0.452$ with the highest Gini coefficient, 0.804.

Table 4 separates concentration from tail steepness. The "$< 1\%$" and "$< 0.1\%$" columns give the aggregate probability mass below each threshold followed by the number of classes in parentheses.

Table 3. Core diversity and imbalance geometry.

Benchmark	Diversity and imbalance metrics
CICIoT2023	$N = 46,686,579$; $K = 8$; $H/H^*/D_1/D_2 = 0.876/0.421/2.40/1.77$; Gini $= 0.765$; max/min $= 2,601\times$
Edge-IIoTset	$N = 2,219,201$; $K = 15$; $H/H^*/D_1/D_2 = 1.209/0.447/3.35/1.86$; Gini $= 0.771$; max/min $= 1,614\times$
ToN-IoT	$N = 461,043$; $K = 10$; $H/H^*/D_1/D_2 = 1.382/0.600/3.98/2.28$; Gini $= 0.584$; max/min $= 288\times$
BoT-IoT	$N = 73,370,443$; $K = 5$; $H/H^*/D_1/D_2 = 0.791/0.491/2.21/2.09$; Gini $= 0.600$; max/min $= 24,280\times$
IoT-23	$N = 6,046,623$; $K = 13$; $H/H^*/D_1/D_2 = 1.159/0.452/3.19/2.59$; Gini $= 0.804$; max/min $= 3,389,036\times$

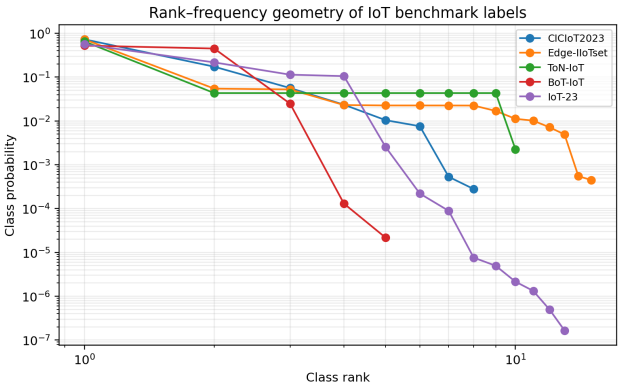


Figure 1. Rank–frequency geometry of the five benchmark label distributions.

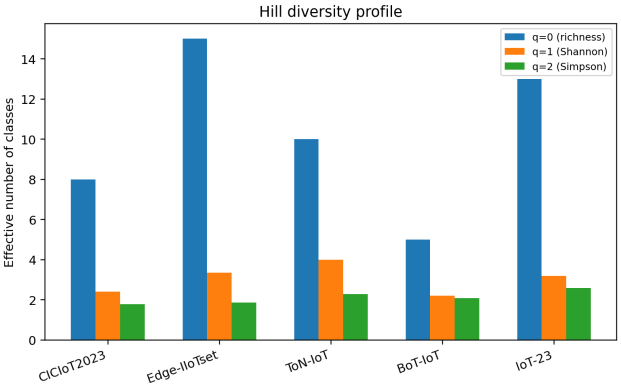


Figure 2. Hill diversity profile for richness ($q = 0$), Shannon diversity ($q = 1$), and Simpson diversity ($q = 2$).

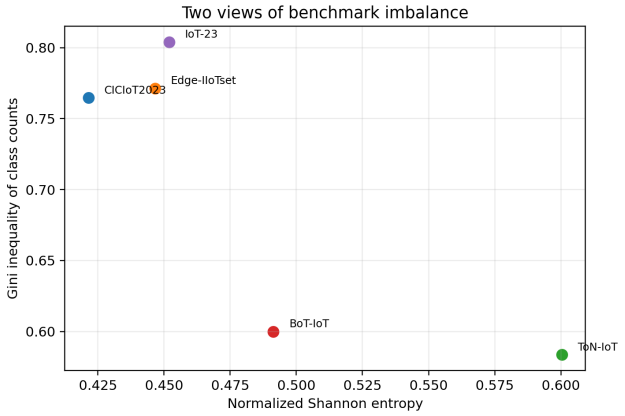


Figure 3. Normalized Shannon entropy versus Gini inequality of class counts.

Table 4. Concentration, entropy deficit, and rare-tail structure.

Benchmark	Concentration and tail metrics
CICIoT2023	HHI $= 0.5638$; KL $= 1.2032$; $\alpha = 3.598$; $R^2 = 0.886$; mass $< 1\%$: 0.841% (3); $< 0.1\%$: 0.081% (2)
Edge-IIoTset	HHI $= 0.5389$; KL $= 1.4989$; $\alpha = 1.985$; $R^2 = 0.759$; mass $< 1\%$: 1.309% (4); $< 0.1\%$: 0.100% (2)
ToN-IoT	HHI $= 0.4385$; KL $= 0.9203$; $\alpha = 1.330$; $R^2 = 0.533$; mass $< 1\%$: 0.226% (1); $< 0.1\%$: 0.000% (0)
BoT-IoT	HHI $= 0.4788$; KL $= 0.8187$; $\alpha = 6.520$; $R^2 = 0.791$; mass $< 1\%$: 0.015% (2); $< 0.1\%$: 0.015% (2)
IoT-23	HHI $= 0.3854$; KL $= 1.4060$; $\alpha = 6.669$; $R^2 = 0.881$; mass $< 1\%$: 0.285% (9); $< 0.1\%$: 0.033% (8)

5. RARE-ATTACK OBSERVABILITY AND EVALUATION RESOLUTION

A class can exist in the parent corpus and disappear from the working sample. If X_i is the number of records from class i in an independent native sample of size m, then X_i is approximately binomial with parameters m and p_i when m is small relative to N; the absence probability and the probability of at least one observation are

$$P(X_i = 0) = (1 - p_i)^m, \qquad P(X_i \geq 1) = 1 - (1 - p_i)^m. \quad (6)$$

Solving the second expression for a 95% observation probability gives

$$m_{i,0.95} = \left\lceil \frac{\ln(0.05)}{\ln(1 - p_i)} \right\rceil. \quad (7)$$

Equation (7) exposes a limit that total-record counts hide. The rarest CICIoT2023 class needs about 10,705 native draws for a 95% chance of being observed at least once; Edge-IIoTset needs about 6,640; ToN-IoT 1,323; and BoT-IoT 138,498. IoT-23 is different. Its singleton label has $p_{\min} = 1/6{,}046{,}623$, so the corresponding $m_{0.95}$ is about 18.1 million under the with-replacement approximation—roughly three times the entire corpus size; without replacement, a 95% chance of including that singleton would require sampling at least 95% of the finite corpus; the approximation is useful here precisely because its impossibility signal is so clear.

Table 5. Native rarest-class observability under random subsampling.

Benchmark	Sampling details: expected count / probability seen	m for 95%
CICIoT2023	$n_{\min} = 13{,}064$, $p_{\min} = 0.00027982$; 0.01%: 1.306 / 72.9%; 0.1%: 13.06 / 100.0%; 1%: 130.64 / 100.0%	10,705
Edge-IIoTset	$n_{\min} = 1{,}001$, $p_{\min} = 0.00045106$; 0.01%: 0.100 / 9.5%; 0.1%: 1.00 / 63.3%; 1%: 10.01 / 100.0%	6,640
ToN-IoT	$n_{\min} = 1{,}043$, $p_{\min} = 0.00226226$; 0.01%: 0.104 / 9.9%; 0.1%: 1.04 / 64.8%; 1%: 10.43 / 100.0%	1,323
BoT-IoT	$n_{\min} = 1{,}587$, $p_{\min} = 0.00002163$; 0.01%: 0.159 / 14.7%; 0.1%: 1.59 / 79.5%; 1%: 15.87 / 100.0%	138,498
IoT-23	$n_{\min} = 1$, $p_{\min} = 0.00000017$; 0.01%: 0.000 / 0.0%; 0.1%: 0.00 / 0.1%; 1%: 0.01 / 1.0%	18,114,063

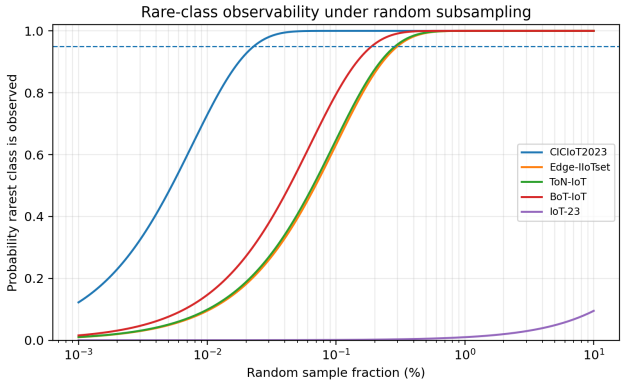


Figure 4. Probability that the rarest class is observed under native random subsampling.

Table 5 evaluates three common subsampling fractions. Each sampling cell is written as "expected rarest-class count / probability of at least one observation." Figure 4 extends the calculation continuously over four orders of magnitude in sampling fraction.

Presence is only the first hurdle. If a class-conditioned metric such as recall is estimated from n independent class records

Table 6. Evaluation-resolution diagnostics from independent class counts.

Benchmark	Evaluation-resolution diagnostics
CICIoT2023	$n_{\min} = 13{,}064$; worst 95% half-width $= 0.009$; classes below 97 / 385 / 25: 0 / 0 / 0
Edge-IIoTset	$n_{\min} = 1{,}001$; worst 95% half-width $= 0.031$; classes below 97 / 385 / 25: 0 / 0 / 0
ToN-IoT	$n_{\min} = 1{,}043$; worst 95% half-width $= 0.030$; classes below 97 / 385 / 25: 0 / 0 / 0
BoT-IoT	$n_{\min} = 1{,}587$; worst 95% half-width $= 0.025$; classes below 97 / 385 / 25: 0 / 0 / 0
IoT-23	$n_{\min} = 1$; worst 95% half-width $= 0.445$; classes below 97 / 385 / 25: 6 / 6 / 4

with observed proportion $\hat{p}$, Wilson's 95% interval has a center and half-width determined by n, $\hat{p}$, and $z = 1.96$ [24]; at the variance-maximizing value $\hat{p} = 0.5$, the half-width is given by Equation (8), which makes the precision available from a finite class count explicit. Precision still depends on n.

$$w(n) = \frac{z}{1 + z^2/n}\sqrt{\frac{0.25}{n} + \frac{z^2}{4n^2}}. \tag{8}$$

This gives $w(97) \approx 0.098$ and $w(385) \approx 0.050$. Table 6 applies the same worst-case calculation to the smallest full-corpus class and counts how many labels fall below the two resolution marks; four benchmarks have at least 1,001 observations in every analyzed label, so their full class populations can in principle support much tighter class-level estimates than IoT-23; six IoT-23 labels have fewer than 97 records, and the same six are below 385. A singleton class has $w \approx 0.445$ at $\hat{p} = 0.5$. The label may be semantically important; statistically it cannot support a precise empirical recall estimate by itself.

These calculations set an upper bound on the evidence available before a train/test split makes a minority class smaller, so they are already favorable to the benchmark: a 20% stratified test split of the Edge-IIoTset fingerprinting class would contain about 200 records, far below the full-class count of 1,001, and a random split of a one-record IoT-23 class cannot place independent observations in both training and test sets. The combinatorial limit remains. No evaluation metric can repair it.

6. MINIMUM-FLATTENING TEMPERED SAMPLING

Training can tolerate a different question. The objective is exposure, and a fixed training budget can be reallocated across classes so that the tail is encountered more often; uniform sampling guarantees equal expected support but can erase much of the native geometry, while native sampling preserves that geometry and may still starve rare labels, which motivates a controlled path between the two endpoints.

$$q_i(\tau) = \frac{p_i^\tau}{Z(\tau)}, \qquad Z(\tau) = \sum_k p_k^\tau, \qquad 0 \le \tau \le 1. \tag{9}$$

At $\tau = 1$, $q = p$, whereas $\tau = 0$ assigns probability $1/K$ to every class. Intermediate values flatten the distribution continuously but retain an explicit relation to the native probabilities, so the amount of prior distortion can be measured exactly and reproduced from the class counts; that property is useful when a researcher wants more rare-class exposure without jumping directly to uniform sampling. The family is a training device. It creates no new observations.

Proposition 1. Entropy decreases monotonically with τ on $[0, 1]$.

Proof. Write $\log q_i = \tau \log p_i - \log Z$. Differentiating $H(q) = -\sum_i q_i \log q_i$ and using $\sum_i q_i' = 0$ gives $H'(\tau) = -\tau \mathrm{Var}_q(\log p)$; the variance is nonnegative, so $H'(\tau) \le 0$ for $\tau \ge 0$, with equality only when the positive p_i are equal or when $\tau = 0$. Lower τ cannot reduce Shannon entropy. □

$$\frac{dH(q(\tau))}{d\tau} = -\tau \mathrm{Var}_q[\ln p_i] \le 0. \tag{10}$$

For any class i, $q_i' = q_i(\ln p_i - \mathbb{E}_q[\ln p])$; when p_i is minimal, its log probability cannot exceed the q-weighted mean, which gives $q_{\min}' \le 0$ and means that lowering τ increases the rarest-class probability while Proposition 1 simultaneously shows that the distribution becomes more entropic. The directions agree. A constrained τ can therefore be chosen without ambiguity inside this one-parameter family.

Given a training budget m and a target expected support r for the rarest class, define

$$\tau^* = \max\left\{\tau \in [0,1] : m \min_i q_i(\tau) \ge r\right\}. \tag{11}$$

Equation (11) chooses the distribution closest to the native endpoint within the tempered family while meeting the support constraint. If the native distribution already meets the target, $\tau^* = 1$. If even uniform allocation gives $m/K < r$, the target is infeasible at that budget; distortion is measured with $D_{\mathrm{KL}}(q\|p)$ and the symmetric Jensen–Shannon divergence, while $\max_i q_i/p_i$ records the largest multiplicative uplift applied to any class.

Table 7 solves Equation (11) at a 1% sampling budget and $r = 25$. Figure 5 varies r, while Figure 6 compares the native and optimized rarest-class support for the $r = 25$ case.

CICIoT2023 needs no adjustment at the 1% budget because its rarest class already contributes an expected 130.64 records. Edge-IIoTset and ToN-IoT need moderate flattening to reach 25, with $\tau^* = 0.855$ and 0.807, while BoT-IoT needs only a small shift, $\tau^* = 0.954$, because its 73.37 million-record total makes a 1% sample large enough for the theft class to start near sixteen expected observations despite the extreme majority-to-minority ratio. IoT-23 behaves differently. Its optimum falls to $\tau^* = 0.450$, $D_{\mathrm{KL}}(q\|p) = 0.219$, and a 2,500-fold probability multiplier for the singleton class; the optimization satisfies the exposure constraint, but the magnitude of the required change is itself evidence that the parent corpus is short of independent tail observations.

Table 7. Minimum-flattening solution at a 1% training budget and target $r = 25$.

Benchmark	Tempered-sampling metrics
CICIoT2023	Native/optimized $E[\min] = 130.64/130.64$; $\tau^* = 1.0000$; $D_1(q) = 2.40$; $D_{KL} = 0.0000$; JS $= 0.0000$; max uplift $= 1.0\times$
Edge-IIoTset	Native/optimized $E[\min] = 10.01/25.00$; $\tau^* = 0.8551$; $D_1(q) = 4.71$; $D_{KL} = 0.0276$; JS $= 0.0066$; max uplift $= 2.5\times$
ToN-IoT	Native/optimized $E[\min] = 10.43/25.00$; $\tau^* = 0.8073$; $D_1(q) = 5.48$; $D_{KL} = 0.0347$; JS $= 0.0085$; max uplift $= 2.4\times$
BoT-IoT	Native/optimized $E[\min] = 15.87/25.00$; $\tau^* = 0.9543$; $D_1(q) = 2.23$; $D_{KL} = 0.0003$; JS $= 0.0001$; max uplift $= 1.6\times$
IoT-23	Native/optimized $E[\min] = 0.01/25.00$; $\tau^* = 0.4504$; $D_1(q) = 4.75$; $D_{KL} = 0.2185$; JS $= 0.0336$; max uplift $= 2{,}500.0\times$

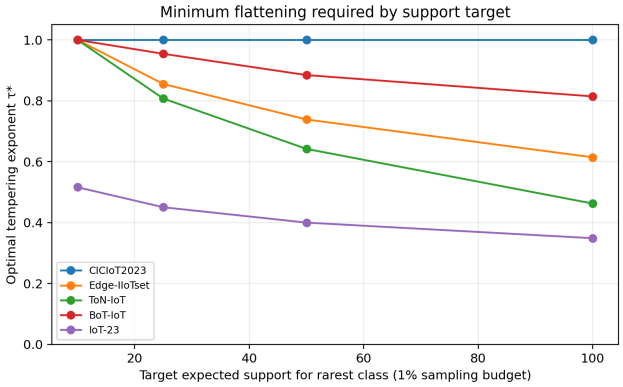


Figure 5. Optimal tempering exponent as the rarest-class support target changes at a 1% training budget.

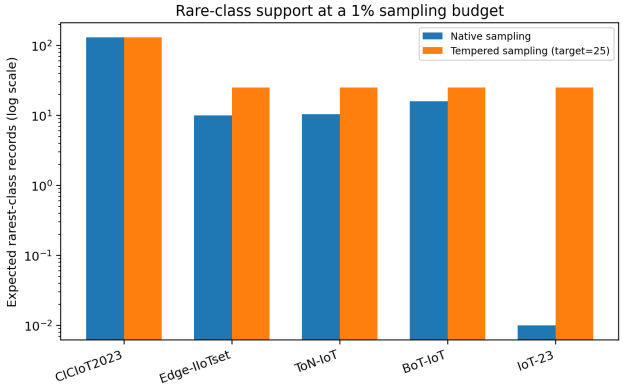


Figure 6. Expected rarest-class support under native and minimum-flattening tempered sampling at a 1% budget.

Repeating one independent IoT-23 observation until its expected training exposure reaches 25 does not create 25 independent examples. It changes the frequency with which one point is presented to a learner; a model may still benefit from class-aware exposure, but evaluation uncertainty is unchanged and memorization risk rises; this is why the 2,500-fold uplift is interpreted as a structural insufficiency indicator and should never be read as a prescription for literal duplication.

Table 8. Sensitivity of τ^* to the target support r at a 1% training budget.

Benchmark	$r = 10$	$r = 25$	$r = 50$	$r = 100$
CICIoT2023	NF	NF	NF	NF
Edge-IIoTset	NF	0.855	0.739	0.614
ToN-IoT	NF	0.807	0.642	0.463
BoT-IoT	NF	0.954	0.884	0.815
IoT-23	0.516	0.450	0.400	0.349

Table 9. Budget sensitivity for target $r = 25$.

Benchmark	0.1% budget	1% budget	5% budget
CICIoT2023	$\tau = 0.911$; KL=0.005	NF	NF
Edge-IIoTset	$\tau = 0.434$; KL=0.517	$\tau = 0.855$; KL=0.028	NF
ToN-IoT	$\tau = 0.198$; KL=0.575	$\tau = 0.807$; KL=0.035	NF
BoT-IoT	$\tau = 0.722$; KL=0.017	$\tau = 0.954$; KL=0.000	NF
IoT-23	$\tau = 0.279$; KL=0.840	$\tau = 0.450$; KL=0.218	$\tau = 0.566$; KL=0.094

7. SENSITIVITY TO SAMPLING BUDGET AND SUPPORT TARGET

A single target can hide the shape of the constraint. Table 8 therefore holds the sampling budget at 1% and raises r from 10 to 100; the entries are τ^*; "NF" means that no tempering is needed because native sampling already meets the target, and "IF" means infeasible even under a uniform distribution at that budget.

The sensitivity curve separates the five benchmarks more clearly than one target value. CICIoT2023 remains at the native distribution through $r = 100$, Edge-IIoTset moves from no change at $r = 10$ to $\tau^* = 0.614$ at $r = 100$, and ToN-IoT begins flattening earlier; IoT-23 stays far below the other curves because several labels have single-digit support, so the tempered family can allocate more training probability to those labels without increasing their empirical diversity. The distinction is structural.

Table 9 changes the budget instead. The support target remains $r = 25$. This separates the effect of label geometry from the effect of computational thrift: a researcher who takes 0.1% of a corpus creates a different observability problem from one who can afford 5%.

At 0.1%, Edge-IIoTset needs $\tau^* = 0.434$ and ToN-IoT $\tau^* = 0.198$ to guarantee 25 expected rare-class records; IoT-23 cannot meet the target because the uniform budget itself is too small. At 5%, the first four benchmarks are native or close

to native, while IoT-23 still requires substantial flattening; increasing m can solve a sampling shortage when the parent corpus contains enough independent minority evidence. It cannot solve a parent-corpus shortage.

8. A QUANTITATIVE AUDIT PROTOCOL FOR IOT BENCHMARK STUDIES

The equations can be reduced to a practical sequence that a model paper can report before training. Step one is to publish the exact class counts after every filtering operation. Percentages alone are insufficient because evaluation resolution depends on n. Step two is to report at least one diversity measure and one tail measure; H^* and D_1 identify how many classes carry effective support, while $n_{\min}$ and the number of labels below 97 or 385 expose the resolution floor; step three is to calculate rare-class observability at the actual subsampling fraction used in the experiment. Step four is to state whether rebalancing changes the training distribution only or also contaminates evaluation; step five is to keep the test set in natural, independent units unless the research question explicitly concerns an artificial prior.

The audit changes how familiar performance numbers should be read. An accuracy of 99.9% on BoT-IoT can coexist with weak evidence about information theft because the two flooding classes dominate the corpus, while a high macro score on IoT-23 can be unstable when a fine label is represented by only a handful of independent records; by contrast, the much smaller processed ToN-IoT network subset can support tighter class-wide estimates when evidence is distributed more evenly. Record count alone cannot settle that comparison.

The same logic extends to federated IoT because Sleem [3] and other federated IDS designs operate under client heterogeneity, where a global class distribution can conceal local absence; if a rare class has probability p globally but is concentrated on one client, the iid observation formula in Equation (6) becomes optimistic for most clients, so a client-level audit should replace p_i with $p_{i,c}$ and report the number of clients on which each class is observed. Non-iid partitioning tightens the requirement.

Explainability has a related dependence: Assudani et al. [2] use SHAP-based interpretation in an IoT IDS, yet feature attributions for a rare class can only reflect the diversity of the rare-class observations that reach the explainer, and oversampling may stabilize optimization while leaving the explanatory sample anchored to duplicated or near-duplicated points. Unique minority counts should accompany attribution summaries.

Real-time detectors add a temporal version of the problem. Salman [4] targets real-time distributed detection, and Abualghanam et al. [6] combine multi-dataset behavior analysis with threat intelligence. A stream can change the effective class prior from hour to hour. Static entropy is then only a baseline. A production audit can compute H_t, $D_{1,t}$, and $q_{\min,t}$ over rolling windows, with change alarms when the label geometry moves beyond a control band; the equations in this paper are therefore compatible with online monitoring even though the empirical study is cross-sectional.

Table 10 summarizes the proposed audit rules. They are reporting requirements and diagnostics. No new classifier is introduced.

8.1 Metric Sensitivity to Rare-Class Improvement

Class geometry also determines how strongly a performance metric reacts when a rare class improves. Let r_i denote recall for class i. In a single-label multiclass problem, the prevalence-weighted recall has the same weighting structure as overall accuracy when each record contributes one correct-or-incorrect outcome, whereas macro recall assigns equal weight to every class; the two summaries are therefore

$$R_{\text{micro}} = \sum_i p_i r_i, \qquad R_{\text{macro}} = \frac{1}{K}\sum_i r_i. \tag{12}$$

If one class improves by Δ while every other class is unchanged, the corresponding score increments are $p_i\Delta$ and Δ/K. Their ratio is independent of Δ,

$$A_i = \frac{(\Delta/K)}{p_i\Delta} = \frac{1}{Kp_i}. \tag{13}$$

Equation (13) is a metric-sensitivity multiplier, not a rule for choosing one score. It measures how much more visible the same class-level improvement becomes under equal class weighting than under native prevalence weighting, and the factor grows as p_i shrinks; in the extreme tail, a stable micro-level aggregate can therefore coexist with a large change in minority-class behavior because the prevalence-weighted metric assigns that class almost no influence. The result is algebraic. It assumes no classifier architecture.

Table 11 evaluates A_i at the rarest label of each benchmark. The contrast spans four orders of magnitude: ToN-IoT has the smallest amplification at about 44×, CICIoT2023 and Edge-IIoTset lie in the hundreds, BoT-IoT exceeds nine thousand, and IoT-23 reaches about 465 thousand because its rarest label has probability 1/6,046,623.

The multiplier also clarifies why metric choice cannot compensate for missing evidence. Macro recall gives a singleton label the same nominal score weight as a majority label, but the singleton estimate still rests on one independent observation; prevalence weighting hides the class, while equal weighting exposes its volatility. A defensible report therefore needs both the metric and the support count. Saito and Rehmsmeier [8] make the corresponding point for precision–recall analysis under imbalance, and the class-count audit supplies the missing denominator for multiclass IoT benchmarks.

A useful reporting pair follows directly: publish macro-oriented performance so that rare classes remain visible, and publish n_i or its test-set counterpart so readers can judge the resolution of each class estimate. Weighted or micro metrics can still be reported because they answer deployment-prevalence questions, but they should not be used alone when the scientific claim is that the detector handles all labeled attack types.

9. INTERPRETATION OF THE FIVE BENCHMARKS

CICIoT2023 is large enough that coarse-category subsampling can be computationally aggressive without immediately

Table 10. Benchmark-adequacy audit rules for IoT security experiments.

Audit item	Status	Reason
Publish exact post-filter class counts	Required	Makes N, p_i, and independent support auditable
Report H^* and D_1	Recommended	Separates label richness from effective diversity
Report $n_{\min}$ and classes below 97/385	Recommended	Shows class-level resolution floor
Compute $P(\text{seen})$ at actual sampling budget	Required when subsampling	Detects silent class omission
Restrict rebalancing to training data	Required for natural-prior evaluation	Prevents synthetic frequency from being mistaken for evidence
Report unique minority records after augmentation	Recommended	Separates exposure from independent information
Use macro/PR-oriented metrics beside accuracy	Recommended	Reduces majority-class masking [7, 8]
Audit client-level counts in federated learning	Required for non-iid claims	Global balance can hide local absence

Table 11. Rarest-class sensitivity of macro and prevalence-weighted recall.

Benchmark	Rarest-class recall sensitivity
CICIoT2023	$K = 8$; $p_{\min} = 0.00027982$; micro weight $= 0.00027982$; macro weight $= 0.12500$; amplification $= 446.71\times$
Edge-IIoTset	$K = 15$; $p_{\min} = 0.00045106$; micro weight $= 0.00045106$; macro weight $= 0.06667$; amplification $= 147.80\times$
ToN-IoT	$K = 10$; $p_{\min} = 0.00226226$; micro weight $= 0.00226226$; macro weight $= 0.10000$; amplification $= 44.20\times$
BoT-IoT	$K = 5$; $p_{\min} = 0.00002163$; micro weight $= 0.00002163$; macro weight $= 0.20000$; amplification $= 9{,}246.43\times$
IoT-23	$K = 13$; $p_{\min} = 0.00000017$; micro weight $= 0.00000017$; macro weight $= 0.07692$; amplification $= 465{,}124.85\times$

losing the rarest category; at 1%, its expected BruteForce support is 130.64 and no tempering is required for the $r = 25$ experiment, yet the weakness lies elsewhere because DDoS alone carries about 72.8% of the category mass, normalized entropy is 0.421, and $D_1 = 2.40$ despite eight categories; a detector can therefore achieve a high micro-averaged score by being excellent on a small effective portion of the label space, which is why class-wise performance must be reported and sample size should not be treated as evidence of balance [12]. The benchmark remains useful.

Edge-IIoTset has fifteen labels, a feature-rich IIoT setting, and a processed population above two million [13, 14]. Its $D_1 = 3.35$ is larger than CICIoT2023's but still far below $K = 15$; the two smallest classes, MITM and Fingerprinting, are near one thousand records; that is enough for a full-population class estimate with a worst-case Wilson half-width around three percentage points, yet ordinary train/validation/test partitioning reduces the independent test support. The optimization result $\tau^* = 0.855$ at a 1% budget quantifies a manageable training-exposure correction that remains within a moderate distortion range.

ToN-IoT is the distributional counterexample to the intuition that bigger is automatically better; the analyzed network subset is only 461,043 records, but nine of ten classes are deliberately substantial and the rare MITM class has 1,043 records [15, 16]; its normalized entropy is 0.600, $D_1 = 3.98$, and Gini=0.584—the best combination in this five-dataset comparison. It still has a majority normal class, so it is not balanced. The point is narrower: support is less concentrated, and class-wide evaluation is correspondingly easier to defend.

BoT-IoT has the largest total population, 73.37 million, and one of the sharpest majority-to-minority ratios at about 24,280:1 [17]. DDoS and DoS together dominate the records. Yet the theft class has 1,587 observations, so the full corpus contains more independent tail evidence than IoT-23's smallest labels; this produces an instructive split between prevalence inequality and absolute support; Gini and H^* say the distribution is strongly concentrated; Wilson resolution says the rarest full-population class is still measurable to roughly ± 2.5 percentage points at the worst-case proportion.

IoT-23 is the case where fine taxonomy meets a sparse tail. Its real malware captures are valuable precisely because they preserve behaviors that a coarser taxonomy might merge [18]. The 6.05-million-record aggregation, however, includes labels with 46, 30, 13, 8, 3, and 1 observations [19]. This drives Gini to 0.804 and makes random subsampling hazardous for those labels. It also explains why $\tau^* = 0.450$ is so extreme. The correct response is to retain or revise the dataset according to the scientific question, instead of automatically discarding the dataset; researchers can merge labels when the scientific question supports that taxonomy, evaluate only classes with sufficient independent support, collect more evidence, or explicitly mark the rarest classes as qualitative cases, with resampling kept separate from claims of statistical resolution.

A comparison of these interpretations shows why one scalar imbalance ratio is inadequate. BoT-IoT has a much larger

max/min ratio than Edge-IIoTset, but its rarest class still has more independent records. IoT-23 has both an extreme ratio and extremely low absolute support. ToN-IoT has the smallest N and the strongest entropy profile. The benchmark audit therefore needs at least three axes: relative geometry, absolute tail counts, and the sampling budget actually used.

10. REPRODUCIBILITY DATA AND FULL SENSITIVITY RESULTS

Table 12 reports every class count used in the calculations at the analytical taxonomy adopted for this study. The counts are taken from the cited benchmark sources [12, 13, 15, 17, 18, 19] and are not reconstructed from plotted percentages; they are the direct inputs to Tables 3–9 and Figures 1–6, so a reader can recompute the diversity, observability, uncertainty, and sampling quantities without access to packet payloads.

The second reproducibility object is the full tempering sensitivity matrix. Table 13 reports τ^* for every combination of sampling fraction and target support used in the sensitivity analysis; "NF" denotes native feasibility and "IF" denotes infeasibility even at $\tau = 0$, which makes the transitions discussed in Section 7 directly auditable and avoids replacing the mathematical result with simulated classifier outcomes.

The matrix also separates a sampling shortage from a parent-corpus shortage. When a target becomes feasible after the budget grows, additional draws can resolve the exposure constraint; when a singleton or near-singleton tail still requires extreme flattening, the limiting quantity is the number of independent observations already present in the benchmark.

11. LIMITATIONS AND REPRODUCIBILITY BOUNDARY

The study deliberately compresses each benchmark to its class-count vector. Feature overlap, temporal leakage, duplicated flows, capture topology, device diversity, and attack realism are outside that vector; a benchmark can pass every distributional check in this paper and still be a poor test of deployment generalization; the proposed audit is therefore necessary for some claims, never sufficient for all of them.

Cross-dataset label semantics are another limit. The eight CICIoT2023 categories, fifteen Edge-IIoTset classes, five BoT-IoT categories, ten ToN-IoT classes, and thirteen IoT-23 labels live at different taxonomic resolutions; entropy can compare their allocation of support because it operates on probabilities; it cannot make the labels semantically equivalent. No conclusion in this paper depends on such equivalence.

The iid observation model in Equation (6) is an approximation. Network flows are temporally and topologically dependent, and random sampling may preserve near-duplicates; dependence generally reduces effective information relative to the nominal count, so the independence calculations can be optimistic; a stronger future audit would estimate an effective sample size from time blocks, devices, sessions, or capture scenarios and replace n_i with the number of independent units relevant to the claim.

The power-tempered family is intentionally narrow. It provides a monotone path and a unique minimum-flattening solution under the rare-class support constraint, but it does not optimize classifier loss, calibration, privacy, energy, or communication cost; other sampling schemes may be better for a particular model; the benefit of Equations (9)–(11) is transparency: the amount of prior distortion is explicit and can be reproduced from counts alone.

The numerical audit is tied to the dataset versions and processed distributions documented in the cited sources. Repositories evolve, preprocessing pipelines change, and a later release can alter the class-count vector even when the dataset name stays the same; any application of this framework should therefore recompute the metrics from the exact files used in the experiment and record those counts with the results. The equations are stable. The inputs are version-specific.

12. CONCLUSION

Five widely used IoT security benchmarks were examined as information distributions before they were treated as machine-learning inputs. The analysis found no simple relation between record count and label adequacy. The 73.37-million-record BoT-IoT corpus has an effective Shannon diversity of 2.21 classes; the 461,043-record ToN-IoT distribution reaches 3.98; CICIoT2023 and Edge-IIoTset contain millions of observations but still place most support in a few classes; IoT-23 goes further: a multi-million-record corpus contains a singleton label, making ordinary random subsampling incapable of reliable rare-label exposure.

The power-tempered sampling family changes training exposure with a controlled amount of flattening. Its entropy obeys $dH/d\tau = -\tau \mathrm{Var}_q(\ln p) \leq 0$, and the rarest-class probability increases as τ falls, so a reproducible τ^* follows from any stated sampling budget and support target; at a 1% budget with $r = 25$, four benchmarks need no to moderate adjustment, whereas IoT-23 requires $\tau^* = 0.450$ and roughly a 2,500-fold uplift of its singleton class. That number marks the limit of rebalancing. Exposure can be changed. Independent evidence cannot.

IoT intrusion-detection papers can therefore strengthen their empirical claims with a short pre-training audit: publish exact class counts, report effective diversity, compute rare-class observability at the chosen sample fraction, quantify class-level evaluation resolution, and separate training rebalancing from test evidence. These checks are inexpensive. They make benchmark-driven intelligent systems easier to compare, reproduce, and interpret.

DATA AVAILABILITY

The study uses class counts reported in the cited public benchmark papers and repositories for CICIoT2023, Edge-IIoTset, ToN-IoT, BoT-IoT, and IoT-23; all derived metrics in the tables follow directly from the equations stated in the manuscript. No private or restricted dataset is used.

REPRODUCIBILITY NOTE

The calculations require only the class-count vectors reported in Table 12 and the equations stated in the manuscript; natural logarithms are used throughout, and probabilities are evalu-

Table 12. Exact class-count vectors used in the analysis.

Benchmark	Label	Count	Share
CICIoT2023	DDoS	33,984,560	72.792997%
CICIoT2023	DoS	8,090,738	17.329901%
CICIoT2023	Mirai	2,634,124	5.642144%
CICIoT2023	Benign	1,098,195	2.352271%
CICIoT2023	Spoofing	486,504	1.042064%
CICIoT2023	Recon	354,565	0.759458%
CICIoT2023	Web	24,829	0.053182%
CICIoT2023	BruteForce	13,064	0.027982%
BoT-IoT	DDoS	38,532,480	52.517715%
BoT-IoT	DoS	33,005,194	44.984319%
BoT-IoT	Normal	9,543	0.013007%
BoT-IoT	Reconnaissance	1,821,639	2.482797%
BoT-IoT	Theft	1,587	0.002163%
IoT-23	Benign	688,812	11.391681%
IoT-23	PartOfAHorizontal PortScan	3,389,036	56.048409%
IoT-23	Okiru	1,313,012	21.714798%
IoT-23	DDoS	638,506	10.559712%
IoT-23	C&C	15,286	0.252802%
IoT-23	C&C-HeartBeat	1,332	0.022029%
IoT-23	Attack	538	0.008898%
IoT-23	C&C-FileDownload	46	0.000761%
IoT-23	C&C-Torii	30	0.000496%
IoT-23	FileDownload	13	0.000215%
IoT-23	C&C-HeartBeat-FileDownload	8	0.000132%
IoT-23	Okiru-Attack	3	0.000050%
IoT-23	C&C-Mirai	1	0.000017%
Edge-IIoTset	Normal	1,615,643	72.802914%
Edge-IIoTset	DDoS_UDP	121,568	5.478008%
Edge-IIoTset	DDoS_ICMP	116,436	5.246753%
Edge-IIoTset	DDoS_TCP	50,062	2.255857%
Edge-IIoTset	DDoS_HTTP	49,911	2.249053%
Edge-IIoTset	SQL_injection	51,203	2.307272%
Edge-IIoTset	Uploading	37,634	1.695836%
Edge-IIoTset	Backdoor	24,862	1.120313%
Edge-IIoTset	Vulnerability_scanner	50,110	2.258020%
Edge-IIoTset	Port_Scanning	22,564	1.016762%
Edge-IIoTset	XSS	15,915	0.717150%
Edge-IIoTset	Ransomware	10,925	0.492294%
Edge-IIoTset	Password	50,153	2.259958%
Edge-IIoTset	MITM	1,214	0.054704%
Edge-IIoTset	Fingerprinting	1,001	0.045106%
ToN-IoT	Normal	300,000	65.069852%
ToN-IoT	Backdoor	20,000	4.337990%
ToN-IoT	DDoS	20,000	4.337990%
ToN-IoT	DoS	20,000	4.337990%
ToN-IoT	Injection	20,000	4.337990%
ToN-IoT	Password	20,000	4.337990%
ToN-IoT	Ransomware	20,000	4.337990%
ToN-IoT	Scanning	20,000	4.337990%
ToN-IoT	XSS	20,000	4.337990%
ToN-IoT	MITM	1,043	0.226226%

Table 13. Minimum-flattening exponent τ^* across budgets and support targets.

Benchmark	Budget	$r = 10$	$r = 25$	$r = 50$	$r = 100$
CICIoT2023	0.1%	NF	0.9105	0.8133	0.7139
CICIoT2023	1.0%	NF	NF	NF	NF
CICIoT2023	5.0%	NF	NF	NF	NF
Edge-IIoTset	0.1%	0.6144	0.4337	0.2805	0.1084
Edge-IIoTset	1.0%	NF	0.8551	0.7387	0.6144
Edge-IIoTset	5.0%	NF	NF	NF	0.8912
ToN-IoT	0.1%	0.4627	0.1983	IF	IF
ToN-IoT	1.0%	NF	0.8073	0.6421	0.4627
ToN-IoT	5.0%	NF	NF	NF	0.8579
BoT-IoT	0.1%	0.8145	0.7217	0.6511	0.5802
BoT-IoT	1.0%	NF	0.9543	0.8845	0.8145
BoT-IoT	5.0%	NF	NF	NF	0.9767
IoT-23	0.1%	0.3488	0.2792	0.2242	0.1654
IoT-23	1.0%	0.5162	0.4504	0.4000	0.3488
IoT-23	5.0%	0.6304	0.5656	0.5162	0.4665

ated without intermediate rounding before derived quantities are calculated. Table 13 supplies the full budget-by-target tempering matrix used for the sensitivity discussion.

REFERENCES

[1] R. Baazeem, "An ensemble boosting algorithm based intrusion detection system for smart internet of things environment," *Journal of Intelligent Systems and Internet of Things*, vol. 13, no. 2, pp. 272–292, 2024.

[2] P. J. Assudani, N. V. S. Pavan Kumar, K. Mohanambal, and R. Chitra, "Explainable artificial intelligence driven intrusion detection system for enhancing reliability and interpretability in IoT based network security solutions," *Journal of Intelligent Systems and Internet of Things*, vol. 17, no. 1, pp. 219–238, 2025.

[3] A. Sleem, "Intelligent and secure detection of cyber-attacks in industrial internet of things: A federated learning framework," *Journal of Intelligent Systems and Internet of Things*, vol. 7, no. 1, pp. 51–61, 2022.

[4] W. A. H. Salman, "A distributed intrusion detection using long short-term memory-gradient repeating unit and enhanced density peak clustering for real-time cyber threat detection," *Journal of Intelligent Systems and Internet of Things*, vol. 18, no. 1, pp. 140–149, 2026.

[5] R. Abdulhammed, S. M. M. Najeeb, R. R. Ali, and M. A. Jubair, "Securing IoT through intrusion detection systems: An overview," *Journal of Intelligent Systems and Internet of Things*, vol. 15, no. 2, pp. 183–190, 2025.

[6] O. Abualghanam, M. Al-Essa, W. Almobaideen, M. Qatawneh, and A. S. Al-Shamayleh, "Proactive

DoS and DDoS attack detection through behavior-based threat intelligence," *Electronics*, vol. 15, no. 12, p. 2559, 2026.

[7] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[8] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, p. e0118432, 2015.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[10] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

[11] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *International Conference on Learning Representations*, 2021.

[12] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment," *Sensors*, vol. 23, no. 13, p. 5941, 2023.

[13] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40 281–40 306, 2022.

[14] A. A. Ahmed and T. A. A. Abdullah, "An enhanced XGBoost-based framework for efficient multi-class cyber threat detection in industrial IoT networks," *Technologies*, vol. 14, no. 5, p. 274, 2026.

[15] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," *IEEE Access*, vol. 8, pp. 165 130–165 150, 2020.

[16] R. Elsayed, R. Hamada, M. Hammoudeh, M. Abdalla, and S. A. Elsaid, "A hierarchical deep learning-based intrusion detection architecture for clustered internet of things," *Journal of Sensor and Actuator Networks*, vol. 12, no. 1, p. 3, 2023.

[17] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

[18] S. Garcia, A. Parmisano, and M. J. Erquiaga, "IoT-23: A labeled dataset with malicious and benign IoT network traffic," Zenodo, 2020.

[19] H. Ghani, S. Salekzamankhani, and B. Virdee, "Statistical and multivariate analysis of the IoT-23 dataset: A comprehensive approach to network traffic pattern discovery," *Journal of Cybersecurity and Privacy*, vol. 5, no. 4, p. 112, 2025.

[20] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, 1948.

[21] M. O. Hill, "Diversity and evenness: A unifying notation and its consequences," *Ecology*, vol. 54, no. 2, pp. 427–432, 1973.

[22] E. H. Simpson, "Measurement of diversity," *Nature*, vol. 163, p. 688, 1949.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.

[24] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927.