Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 11 Issue 42 (March 2026) PP. 337-353

JOURNAL OF INFORMATION SYSTEM AND TECHNOLOGY MANAGEMENT (JISTM)
www.gaexcellence.com/jistm

BENIGN-AWARE HISTOGRAM GRADIENT BOOSTING FOR MALICIOUS IOT NETWORK TRAFFIC DETECTION

Mohd Noor Derahman[1*], Qin Zezheng[2], Azizol Abdullah[3], Shafinah Kamarudin[4]

[1] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
mnoord@upm.edu.my https://orcid.org/0000-0003-1120-6728
[2] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
GS68378@upm.edu.my https://orcid.org/0000-0001-5749-3449
[3] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
azizol@upm.edu.my https://orcid.org/0000-0001-8321-9259
[4] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
shafinah@upm.edu.my https://orcid.org/0000-0002-5705-9172
*Corresponding Author

Article Info:

Article history:
Received date: 29.01.2026
Revised date: 22.02.2036
Accepted date: 18.03.2026
Published date: 30.03.2026

To cite this document:

Derahman, M. N., Qin, Z., Abdullah, A., & Kamarudin, S. (2026). Benign-Aware Histogram Gradient Boosting for Malicious Iot Network Traffic Detection. *Journal of Information System and Technology Management, 11* (42), 337-353.

Abstract:

Detecting malicious traffic in Internet of Things (IoT) networks remains challenging because flow distributions are highly skewed, attack behaviours evolve quickly, and practical deployments must balance accuracy with computational cost. This study evaluates five classical machine learning models on IoT-23 and CICIoT2023 under multiple sample sizes and preprocessing settings. The experimental design includes 1,000, 5,000, 10,000, and 50,000-record subsets, median imputation, five-fold stratified cross-validation, explicit hyperparameter tuning, SMOTE-based imbalance analysis, and training and inference cost measurement. In addition to the five baseline models, the study introduces a benign-aware histogram gradient boosting variant (BA-HGB) that applies tuned cost- sensitive weighting to the minority benign class without synthetic data generation. On CICIoT2023, BA-HGB achieved the best five-fold macro- F1 score relative to the baseline models on the 10,000-sample benchmark (0.8898 +/- 0.0153), the best macro-F1 at 50,000 samples (0.8996 +/- 0.0038), and the highest ROC-AUC (0.9971 +/- 0.0003). An ablation in- side the HGB family further showed that all HGB variants outperformed the RF and GB baselines, whereas SMOTE consistently reduced both macro-F1 and benign-class F1. These results support the generalizability of the findings and show that histogram-based boosting is a strong practical direction for IoT intrusion detection, while imbalance handling mainly changes the accuracy-stability trade-off within that family.

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353

DOI: 10.35631/JISTM.1142020 **Keyword:**

CICIoT2023; Class Imbalance; Histogram Gradient Boosting; IoT Security; Machine Learning; Malicious Traffic Detection

Introduction

The rapid growth of Internet of Things (IoT) deployments has expanded the attack surface of modern networks and increased the operational importance of detecting malicious traffic accurately and efficiently (Shafiq et al., 2022; Zhang & Lazaro, 2024; Abdullah et al., 2024; Inuwa & Das, 2024). Classical machine learning remains attractive in this setting because it can be trained on tabular flow features, deployed with limited hardware, and interpreted more easily than many deep models (Alzaabi and Mehmood, 2024; Kruschel et al., 2025; Wigerthale and Reich, 2024). At the same time, recent deep and privacy-preserving IoT security systems show that stronger detection performance often comes with greater implementation complexity and deployment cost (Kamal and Mashaly, 2025; Khan, 2024). However, practical conclusions are only trustworthy when they are tested under multiple data conditions, validated on more than one dataset, and discussed in terms of both effectiveness and computational cost.

Comparative studies on IoT malicious-traffic detection often report strong performance for ensemble methods, yet many remain limited to a single dataset, narrow sample-size set- tings, under-justified preprocessing, and little discussion of imbalance handling or deployment cost. These limitations make it difficult to judge whether observed model rankings are robust, generalizable, and practically useful for deployment.

This study therefore combines the IoT-23 benchmark with a second dataset, CI- CIoT2023, to test whether the same ordering of models remains valid under a broader IoT attack space. We use the public CICIoT2023 release because it contains diverse attacks collected from real devices and therefore provides a stronger basis for cross-dataset validation (Neto et al., 2023). To keep the detection task aligned across datasets, all non-benign CICIoT2023 labels are mapped to a single malicious class, while the full attack diversity is preserved in the source data. In addition, the study introduces a benign-aware histogram gradient boosting variant designed specifically for the asymmetric benign-versus-malicious balance observed in the CICIoT2023 subset used in this study.

This study makes four contributions. First, it evaluates 1,000, 5,000, 10,000, and 50,000 samples on CICIoT2023 to provide a broader view of scale sensitivity. Second, it strengthens methodological credibility through median imputation, stratified five-fold cross-validation, explicit hyperparameter tuning, and SMOTE-based imbalance handling. Third, it proposes a benign-aware histogram gradient boosting variant (BA-HGB) that uses tuned class weighting

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353
to improve minority benign-class recovery without synthetic traffic generation. Fourth, it discusses practical deployment trade-offs by reporting training time and inference latency in addition to predictive quality.

Literature Review

Classical Machine Learning for IoT Traffic Detection

Machine learning has become a standard approach for malicious traffic detection because manually crafted signatures struggle to keep pace with evolving attack behaviors (Shafi et al., 2022; Zhang & Lazaro, 2024; Abdullah et al., 2024; Ness et al., 2025). Logistic Regression and linear Support Vector Machines are often used as interpretable baselines; K-Nearest Neighbors provides a non-parametric alternative for local pattern matching; and tree-based ensembles such as Random Forest and Gradient Boosting are regularly reported as strong per- formers on structured security data (Jabardi, 2025; Açıkkar and Tokgöz, 2025; Malalha et al., 2024; Rimal et al., 2025; Abolmaali et al., 2024; Inuwa & Das, 2024; Imani et al., 2025). The common pattern across these studies is that ensemble models can capture non-linear feature interactions more effectively than linear baselines, but they may impose higher training cost and reduced interpretability (Kruschel et al., 2025; Mohale and Obagbuwa, 2025; Wiggerthale and Reich, 2024).

Remaining Research Gaps

Despite the maturity of classical algorithms, three gaps remain common in comparative studies. First, results are often reported on a single dataset, which makes it difficult to assess whether findings are dataset specific. Second, preprocessing choices such as feature scaling, missing-value imputation, feature construction, and class balancing are sometimes under-justified even though they strongly affect distance-based and margin-based models (Su- jon et al., 2024; Imani et al., 2025; Bala and Behal, 2024; AlSalehy and Bailey, 2025; Zhang et al., 2024). Third, many papers emphasize predictive accuracy while giving limited attention to hyperparameter sensitivity, validation stability, runtime cost, and the evaluation of minority classes under imbalance (Rimal et al., 2024; Diallo et al., 2024; Li, 2024; Zhu et al., 2024). The methodology adopted in this paper is designed to address those gaps directly.

Methodology

Datasets and Problem Definition

The first dataset used in this paper is IoT-23, a labelled IoT network-traffic dataset collected from real devices and malware scenarios by the Stratosphere Laboratory (Garcia et al., 2020). To improve generalizability, we additionally evaluate the same five models on CI- CIoT2023 (Neto et al., 2023). The public CICIoT2023 subset used in this study contains 712,311 labelled records, 39 numeric features, 33 attack labels, and 16,577 benign samples. Table 1 summarizes its composition.

Both datasets are used for binary malicious-traffic detection. For CICIoT2023, the original labels are converted into a binary target where BENIGN is mapped to the benign class and all attack labels are mapped to the malicious class. This choice keeps the task definition consistent across datasets while still testing the models on a more diverse set of underlying attacks.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353

Table 1: Summary Of the CICIoT2023 Subset Used in This Study

Statistic	Value
Total records	712,311
Feature count	39
Missing values	22
Attack labels	33
Malicious samples	695,734
Benign samples	16,577
Benign ratio	2.33%

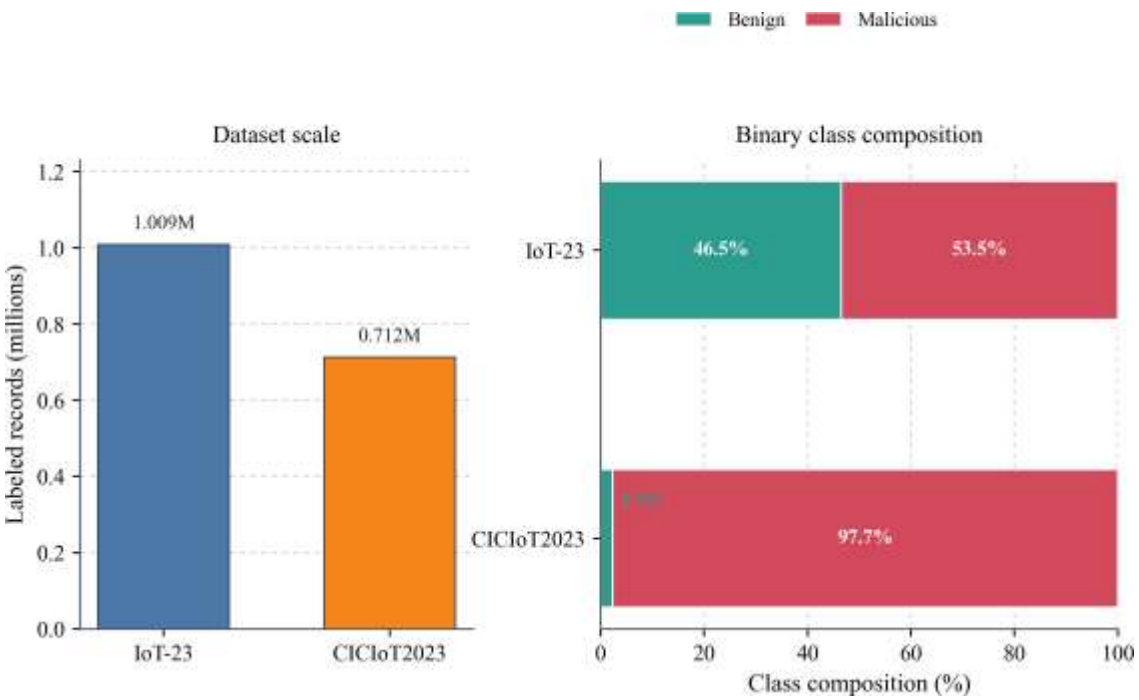

Figure 1: Scale And Binary Class-Composition Comparison Between IoT-23 And CICIoT2023

Figure 1 complements Table 1 by contrasting the two datasets directly. IoT-23 contains 1,008,748 labeled flows and is relatively balanced, with 46.52% benign traffic and 53.48% malicious traffic. By contrast, the CICIoT2023 subset used in this study contains 712,311 records but only 2.33% of benign traffic after collapsing the 33 attack labels into a single malicious class. This much stronger skew explains why the experiments place greater emphasis on macro- F1, stratified validation, and explicit imbalance handling.

Preprocessing and Experimental Design

The preprocessing pipeline is designed to improve methodological rigor and reduce avoidable bias. Missing values are handled with median imputation rather than zero-filling. This decision preserves the empirical distribution of numeric flow features more faithfully and reduces the risk of injecting artificial low-activity patterns into the data, which is consistent with broader preprocessing and data-quality guidance in recent machine learning studies (Bala and Behal, 2024; AlSalehy and Bailey, 2025; Zhang et al., 2024). Standardization is applied to Logistic Regression, KNN, and SVM because these models are sensitive to feature scale, while tree-based

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353

models are trained without scaling, in line with common practice (Sujon et al., 2024). The IoT-23 benchmark uses four conditions: 1,000 and 5,000 samples, each with and without standardization. The CICIoT2023 experiments extend the scale analysis to 1,000, 5,000, 10,000, and 50,000 samples. For each sample size, a stratified three-fold validation is performed to reduce the variance associated with any single train-test split. In addition, a five-fold stratified cross-validation is performed on a 10,000-sample subset to provide a direct estimate of validation stability.

Cross-dataset context is summarized in Figure 1, while exploratory analysis of the IoT- 23 benchmark is retained in Table 3 and Figure 2. These descriptive summaries highlight sparse flows, strong class-imbalance differences across datasets, and correlated numeric features, which motivated the preprocessing and validation choices.

Models, Hyperparameter Tuning, and Evaluation Metrics

The benchmark comparison considers five classical baseline models: Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB). To move beyond a pure benchmarking paper, we additionally introduce a benign-aware histogram gradient boosting variant (BA-HGB). BA-HGB uses histogram-based gradient boosting and assigns a tuned training weight to the minority benign class so that benign-flow recall can improve without generating synthetic samples. Hyperparameters are tuned with grid search on a 10,000-sample stratified subset of CICIoT2023 using macro-F1 as the selection criterion. For BA-HGB, the search covers learning rate, tree depth, boosting iterations, minimum leaf size, and the benign-class weight. This design follows the broader literature showing that careful hyperparameter search and task-specific evaluation choices materially affect model ranking under class imbalance (Rimal et al., 2024; Diallo et al., 2024; Zhu et al., 2024). Macro-F1 is emphasized because it treats the benign and malicious classes equally and therefore provides a more meaningful measure under class imbalance than accuracy alone.

The tuned settings are shown in Table 2. These values are then reused in the sample- size, cross-validation, imbalance, timing, and focused BA-HGB comparison experiments to keep the comparisons consistent. Accuracy and ROC-AUC are also reported to support com- parison with prior intrusion-detection studies (Li, 2024; Rimal et al., 2025).

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 11 Issue 42 (March 2026) PP. 337-353

Table 2: Selected Hyperparameters After Grid Search on CICIoT2023

Model	Selected hyperparameters	CV macro-F1
Logistic Regression	C=1.0	0.8633
KNN	n_neighbors=11, weights=distance	0.8433
SVM	C=0.1	0.8600
Random Forest	max_depth=None, min_samples_split=5, n_estimators=200	0.8918
Gradient Boosting	learning_rate=0.1, max_depth=3, n_estimators=100	0.8973
BA-HGB	learning_rate=0.05, max_depth=8, max_iter=300, min_samples_leaf=20, benign_weight=2.0	0.8889

Class Imbalance and Computational Cost

Class imbalance is handled explicitly in experiments. Because benign traffic is the minority class in the CICIoT2023 subset used in this study, SMOTE is applied inside the training folds only, so that synthetic sampling does not leak information into validation data. The effect of SMOTE is evaluated by comparing macro-F1 and benign-class F1 with and without oversampling. The proposed BA-HGB variant addresses the same imbalance from a complementary angle by using cost-sensitive weighting rather than synthetic resampling, and an ablation study later compares no weighting, moderate weighting, stronger weighting, and SMOTE within the same HGB family. This framing is consistent with recent comparative work showing that resampling and cost-sensitive strategies can behave very differently across model families and imbalance regimes (Imani et al., 2025; Zhu et al., 2024; Diallo et al., 2024). To make the paper more useful for deployment decisions, we also record training time and average inference time per sample on a 50,000-sample subset.

Results and Discussion

IoT-23 Benchmark

The IoT-23 benchmark supports the conclusion that ensemble models remain the most reliable classical options for malicious traffic detection. Figure 3 shows that RF and GB consistently match or exceed the baselines across the four experimental conditions. Figure 4 leads to the same interpretation from a threshold-independent perspective, with the ensemble models providing the most favorable ROC curves and the highest AUC values.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353

These results remain important because they isolate the effects of sample size and standardization on a widely used malware-traffic benchmark. They are complemented by the CI- CIoT2023 analysis, which tests whether the same conclusions hold under broader attack diversity and more rigorous validation.

Table 3 shows that IoT-23 is dominated by sparse and short flows: the median duration is zero, the median number of original packets is only one, and both response-packet measures have medians of zero. At the same time, the large gap between the median and the maxi- mum values confirm a long-tailed distribution in traffic volume, so a small fraction of flows is much heavier than the typical observation. Figure 2 complements this pattern by showing clear positive associations among packet and byte variables, especially between orig_pkts, orig_ip bytes, resp_pkts, and resp_ip bytes. These dependencies help explain why tree-based ensembles remain strong in this setting: they can model correlated and non-linear traffic signals without requiring the stronger linearity assumptions imposed by simpler baselines.

Table 3: Summary Statistics of Selected IoT-23 Numeric Features Used in The IoT-23 Benchmark

Statistic	duration	orig pkts	orig ip bytes	resp pkts	resp ip bytes
Count	1,008,749	1,008,749	1,008,749	1,008,749	1,008,749
Mean	0.6748	1.4962	81.1455	0.1425	9.0492
Std	2.3356	1.7412	94.7309	1.8504	119.6775
Min	0	0	0	0	0
25%	0	1	40	0	0
50%	0	1	60	0	0
75%	0	1	60	0	0
Max	293.0102	60	2,990	75	9,415

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353

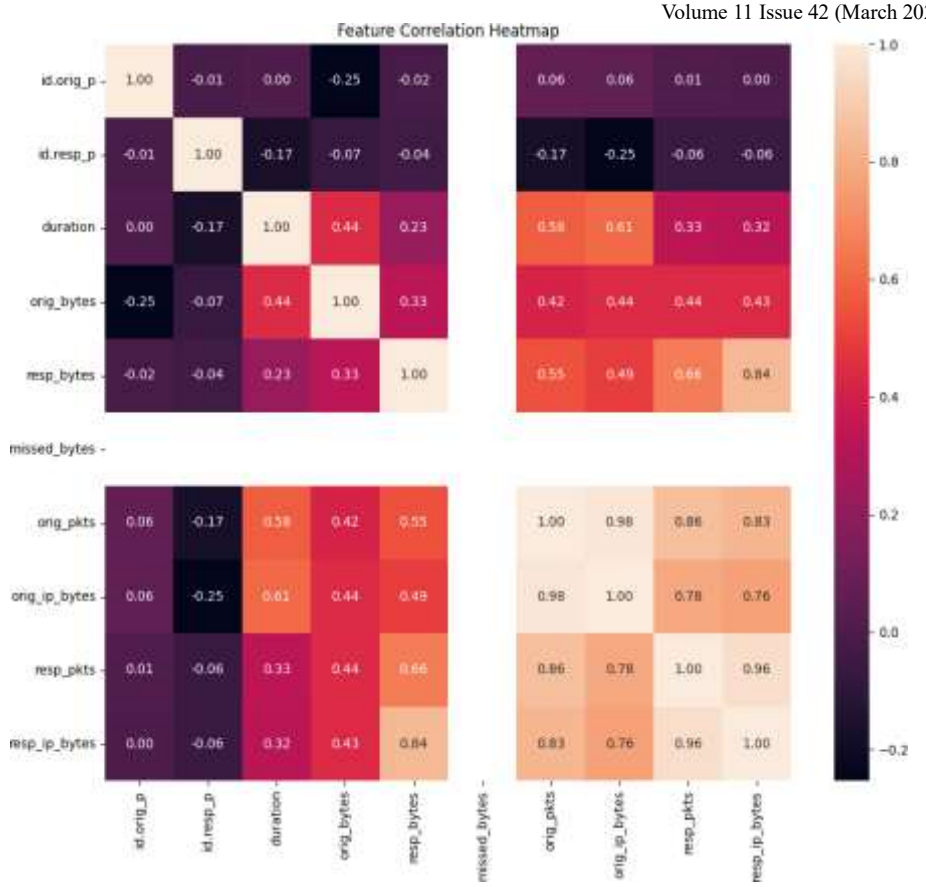

Figure 2: Correlation Heatmap of The IoT-23 Numeric Features

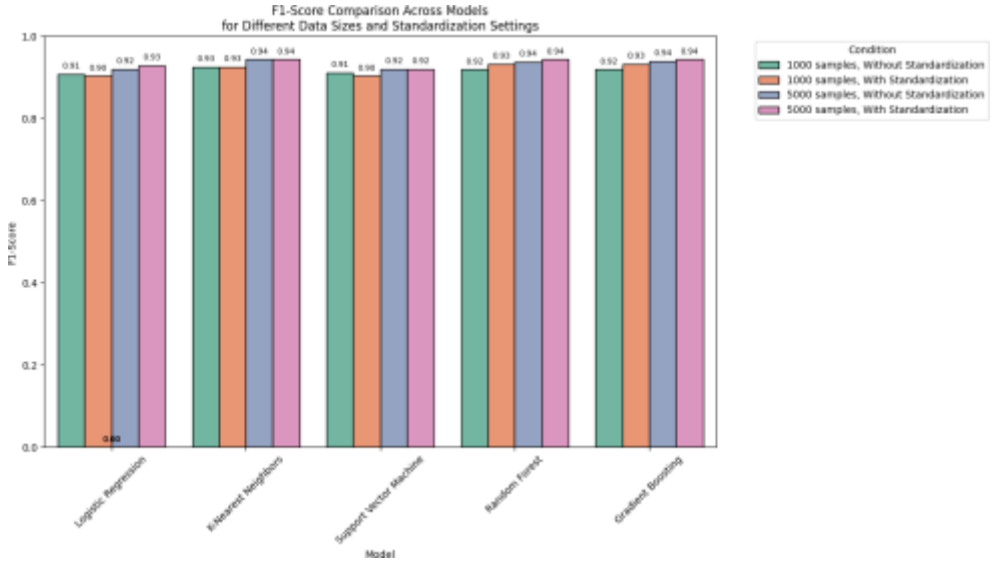

Figure 3: Model Comparison on IoT-23 Under Four Experimental Conditions

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353

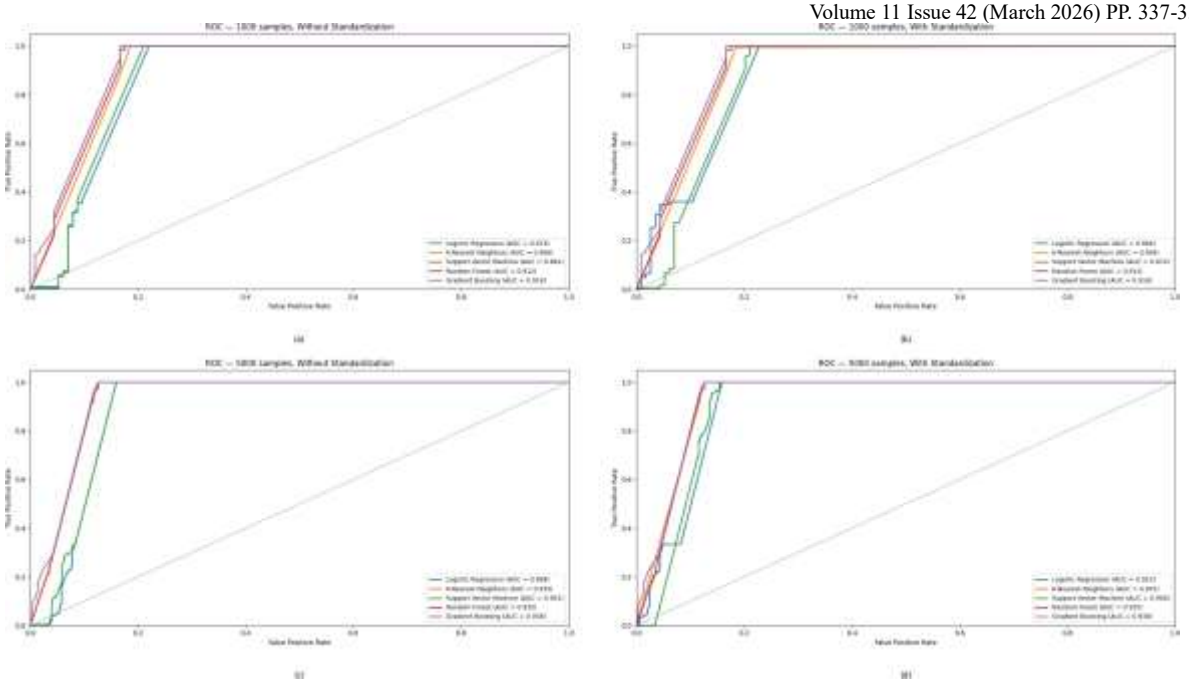

Figure 4: ROC Curves for The IoT-23 Benchmark

Cross-Dataset Validation on CICIoT2023

The new CICIoT2023 experiments confirm the main ranking observed on IoT-23 while adding a stronger test of generalizability. Tables 4 and 5, together with Figure 5, show that RF and GB remain the most competitive models across all four sample sizes. At 50,000 samples, RF achieves the best macro-F1 score, 0.8923 +/- 0.0097, followed by GB at 0.8843 +/- 0.0096. The stronger performance of the ensemble models is therefore not confined to a single dataset.

The expanded scale analysis addresses the limitation of evaluating only two sample sizes. The results are not strictly monotonic for every model because larger subsets expose more heterogeneous traffic patterns and therefore constitute a harder prediction task. However, the ranking is stable: RF and GB remain the top-performing models, while LR, SVM, and KNN form a second tier with lower macro-F1 values. This stability is more informative than a simple monotonic increase because it suggests that the conclusion is robust to dataset size and attack diversity.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353

Table 4: CICIoT2023 Performance For 1,000 And 5,000 Samples

Sample size	Model	Accuracy	Macro-F1	ROC-AUC
1000	Random Forest	0.9890 ± 0.0037	0.8845 ± 0.0408	0.9951 ± 0.0014
1000	Logistic Regression	0.9860 ± 0.0037	0.8468 ± 0.0251	0.9048 ± 0.0613
1000	SVM	0.9860 ± 0.0037	0.8468 ± 0.0251	0.8687 ± 0.0959
1000	Gradient Boosting	0.9860 ± 0.0086	0.8345 ± 0.1053	0.9936 ± 0.0029
1000	KNN	0.9850 ± 0.0065	0.8142 ± 0.0891	0.9941 ± 0.0023
5000	Gradient Boosting	0.9872 ± 0.0012	0.8575 ± 0.0074	0.9955 ± 0.0012
5000	Random Forest	0.9872 ± 0.0003	0.8511 ± 0.0032	0.9952 ± 0.0005
5000	SVM	0.9838 ± 0.0024	0.8254 ± 0.0162	0.9850 ± 0.0124
5000	Logistic Regression	0.9838 ± 0.0030	0.8245 ± 0.0321	0.9850 ± 0.0122
5000	KNN	0.9810 ± 0.0033	0.7882 ± 0.0322	0.9917 ± 0.0019

Table 5: CICIoT2023 Performance For 10,000 And 50,000 Samples

Sample size	Model	Accuracy	Macro-F1	ROC-AUC
10000	Random Forest	0.9898 ± 0.0007	0.8862 ± 0.0058	0.9965 ± 0.0008
10000	Gradient Boosting	0.9892 ± 0.0010	0.8834 ± 0.0138	0.9965 ± 0.0004
10000	SVM	0.9867 ± 0.0012	0.8561 ± 0.0074	0.9848 ± 0.0139
10000	Logistic Regression	0.9865 ± 0.0009	0.8514 ± 0.0053	0.9876 ± 0.0104
10000	KNN	0.9846 ± 0.0018	0.8444 ± 0.0121	0.9894 ± 0.0068
50000	Random Forest	0.9903 ± 0.0008	0.8923 ± 0.0097	0.9969 ± 0.0005
50000	Gradient Boosting	0.9895 ± 0.0008	0.8843 ± 0.0096	0.9967 ± 0.0004
50000	Logistic Regression	0.9867 ± 0.0009	0.8507 ± 0.0116	0.9939 ± 0.0008
50000	KNN	0.9860 ± 0.0006	0.8507 ± 0.0028	0.9905 ± 0.0030
50000	SVM	0.9865 ± 0.0006	0.8483 ± 0.0078	0.9929 ± 0.0002

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353

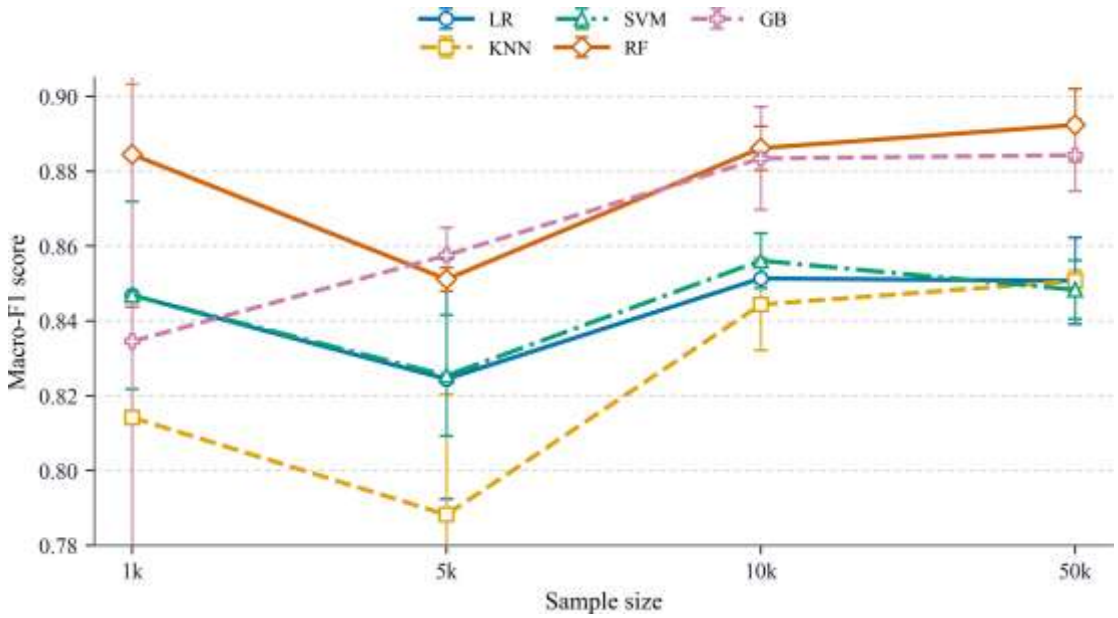

Figure 5: Macro-F1 Across Four CICIoT2023 Sample Sizes (Three-Fold Means with Standard-Deviation Error Bars)

Validation Stability, Imbalance Handling, and Practical Implications

The five-fold stratified cross-validation results in Table 6 further supporting the same conclusion. GB attains the best mean macro-F1, 0.8772 +/- 0.0173, while RF records the highest ROC-AUC, 0.9965 +/- 0.0012. The low standard deviations of the ensemble models indicate that their advantage is not caused by a favorable split. In contrast, the non-ensemble models show both lower macro-F1 and larger variance, especially KNN.

The imbalance analysis reveals a more nuanced finding. Table 7 shows that SMOTE improves both macro-F1 and benign-class F1 for RF and GB, but reduces performance for LR, SVM, and KNN. This means that handling imbalance is necessary, but the treatment should be matched to the learning algorithm rather than applied blindly. For the best-performing models in this study, oversampling is beneficial; for the weaker baselines, it introduces noise that outweighs the benefit of rebalancing.

Table 8 adds an explicit deployment perspective. KNN has the lowest apparent training cost because it mainly stores the training data, but it has the slowest inference among the five models. GB delivers the strongest average validation quality, yet it also requires the longest training time, 7.269 seconds, on the 50,000-sample benchmark. RF therefore offers a practical compromise in this study: it is substantially faster to train than GB while maintaining the best large-sample macro-F1 and the strongest ROC-AUC.

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 11 Issue 42 (March 2026) PP. 337-353

Table 6: Five-Fold Stratified Cross-Validation Results on A 10,000-Sample CICIoT2023 Subset

Model	Accuracy	Macro-F1	ROC-AUC
Gradient Boosting	0.9887 ± 0.0019	0.8772 ± 0.0173	0.9963 ± 0.0011
Random Forest	0.9888 ± 0.0013	0.8750 ± 0.0118	0.9965 ± 0.0012
SVM	0.9870 ± 0.0025	0.8607 ± 0.0200	0.9858 ± 0.0179
Logistic Regression	0.9867 ± 0.0021	0.8547 ± 0.0174	0.9870 ± 0.0157
KNN	0.9853 ± 0.0031	0.8489 ± 0.0281	0.9898 ± 0.0094

Table 7: Effect Of SMOTE On CICIoT2023

Model	Macro-F1	Macro-F1 with SMOTE	Benign F1	Benign F1 with SMOTE
Random Forest	0.8750	0.8956	0.7558	0.7968
Gradient Boosting	0.8772	0.8854	0.7602	0.7772
KNN	0.8489	0.8449	0.7054	0.7001
Logistic Regression	0.8547	0.8430	0.7162	0.6964
SVM	0.8607	0.8362	0.7280	0.6835

Table 8: Training And Inference Cost on A 50,000-Sample CICIoT2023 Subset

Model	Training time (s)	Inference time (ms/sample)
KNN	0.136	0.0597
SVM	0.676	0.0023
Random Forest	0.844	0.0521
Logistic Regression	1.351	0.0029
Gradient Boosting	7.269	0.0027

Proposed Benign-Aware Histogram Gradient Boosting

The baseline comparison above establishes the strongest existing classical baselines, but a practical contribution also requires a task-adapted improvement rather than another ranking exercise. To address that need, we designed a benign-aware histogram gradient boosting variant (BA-HGB) that keeps the same tabular intrusion-detection setting but modifies the training

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353

objective with a tuned 2.0x benign-class weight. This small change is important in practice because the CICIoT2023 subset used in this study is heavily dominated by malicious samples, so a standard loss function tends to favor the majority class too strongly.

Table 9: Focused Comparison Between the Proposed BA-HGB Variant and The Strongest Baseline Ensembles

Model	10k macro-F1	10k benign F1	50k macro-F1	50k ROC-AUC
Random Forest	0.8750 ± 0.0118	0.7558 ± 0.0230	0.8923 ± 0.0097	0.9969 ± 0.0005
Gradient Boosting	0.8772 ± 0.0173	0.7602 ± 0.0336	0.8843 ± 0.0096	0.9967 ± 0.0004
BA-HGB (proposed)	0.8898 ± 0.0153	0.7850 ± 0.0297	0.8996 ± 0.0038	0.9971 ± 0.0003

Table 9 shows that BA-HGB achieves the best performance at both scales. On the 10,000-sample five-fold benchmark, it reaches a macro-F1 score of 0.8898 +/- 0.0153 and a benign-class F1 score of 0.7850 +/- 0.0297, exceeding RF by 1.48 and 2.93 percentage points, respectively. On the 50,000-sample three-fold benchmark, BA-HGB reaches a macro- F1 score of 0.8996 +/- 0.0038 and an ROC-AUC of 0.9971 +/- 0.0003, again outperforming both RF and GB. The gain is therefore not merely a marginal reordering of existing baselines; it shows that the histogram-boosting family is especially competitive in this IoT intrusion- detection setting.

The proposed method also scales well as more data become available. Across 1,000, 5,000, 10,000, and 50,000 samples, BA-HGB increases macro-F1 from 0.8363 to 0.8996 and benign-class F1 from 0.6802 to 0.8043. On the 50,000-sample focused comparison benchmark, BA-HGB trains in 0.606 seconds while maintaining an inference latency of 0.0040 milliseconds per sample. A permutation-importance analysis further shows that Number is the dominant feature for BA-HGB (importance 0.3930), followed by IAT (0.0781), HTTPS (0.0679), and Max (0.0656). This ranking suggests that packet-count regularity, inter-arrival timing, protocol context, and burst size are the main signals exploited by the proposed model, which adds a more interpretable explanation of why the method improves benign-traffic recovery.

Table 10: BA-HGB Ablation Across Four Imbalance-Handling Strategies

Variant	10k macro-F1	10k benign F1	50k macro-F1	50k benign F1
No weighting	0.8949 ± 0.0112	0.7949 ± 0.0218	0.9003 ± 0.0080	0.8053 ± 0.0158
Weight = 1.5	0.8930 ± 0.0163	0.7911 ± 0.0317	0.8992 ± 0.0060	0.8034 ± 0.0117
Weight = 2.0	0.8898 ± 0.0153	0.7850 ± 0.0297	0.8996 ± 0.0038	0.8043 ± 0.0075
SMOTE	0.8835 ± 0.0186	0.7731 ± 0.0360	0.8864 ± 0.0054	0.7785 ± 0.0106

Table 10 separates the contribution of the histogram-boosting backbone from the choice of imbalance treatment. The strongest mean performance is obtained by the unweighted HGB variant, which reaches 0.8949 macro-F1 at 10,000 samples and 0.9003 macro-F1 at 50,000 samples. The tuned 2.0x benign-weight configuration used in the focused comparison remains highly competitive and yields the lowest large-sample variance (0.0038), making it the most stable of the weighted settings on the 50,000-sample benchmark. In contrast, SMOTE is consistently the weakest HGB variant, reducing both macro-F1 and benign-class F1 at both

scales. This ablation suggests that the main gain comes from the histogram-based boosting architecture itself, while weighting acts as a secondary stability-oriented adjustment and synthetic oversampling is not beneficial for this model family.

Limitations

This study still has several limitations that should be acknowledged clearly. First, the CICIoT2023 analysis uses a public subset rather than the full release, so the new results should be interpreted as strong cross-dataset evidence rather than a complete dataset-wide benchmark. Second, the task is framed as binary malicious-traffic detection; a full multi-class analysis of the 33 attack labels would be a useful extension but would answer a slightly different research question. Third, BA-HGB is a task-specific adaptation of histogram boosting rather than a wholly new learning family, so its contribution is practical rather than theoretical. Fourth, the paper remains focused on classical machine learning, so it does not claim superiority over recent deep, privacy-preserving, or robustness-oriented alternatives (Kamal and Mashaly, 2025; Khan, 2024; Eleftheriadis et al., 2024).

Conclusion

This study strengthens the evidence for classical IoT malicious-traffic detection by combining IoT-23 and CICIoT2023, broader sample-size analysis, explicit hyperparameter tuning, five-fold stratified cross-validation, median imputation, SMOTE-based imbalance handling, runtime reporting, and a new benign-aware histogram gradient boosting variant. Across both datasets, the main conclusion remains consistent: tree-based ensembles are the strongest classical family for malicious IoT traffic detection. More importantly, the proposed BA-HGB variant improves on the strongest baseline ensembles by combining higher macro-F1, stronger benign-class recovery, fast training, slightly better ROC-AUC on CICIoT2023, and a clearer feature-level interpretation of the decision process. The experiments therefore provide a more defensible, task-adapted basis for model selection.

Acknowledgements:	The authors would like to express their sincere gratitude to Fakulti Sains Komputer dan Teknologi Maklumat, Universiti Putra Malaysia for providing the necessary resources and support throughout the course of this research. Special appreciation is extended to colleagues and peers who contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper.
Funding Statement:	No Funding
Conflict of Interest Statement:	The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have contributed to this work and approved the final version of the manuscript for submission to the Journal of Information System and Technology Management (JISTM)

Ethics Statement:	This study did not involve any human participants, animals, or sensitive data requiring ethical approval. The authors confirm that the research was conducted in accordance with accepted academic integrity and ethical publishing standards.
Author Contribution Statement:	All authors contributed significantly to the development of this manuscript. Mohd Noor Derahman was responsible for the conceptualization, methodology, and overall supervision of the study. Qin Zezheng handled data collection, analysis, and interpretation of results. Azizol Abdullah contributed to drafting, and critical revision of the manuscript. Shafinah Kamarudin contributed to drafting, and critical revision of the manuscript. All authors read and approved the final version of the manuscript prior to submission.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 42 (March 2026) PP. 337-353

References

Abdullah, M. M., Khan, H., Farhan, M., Khadim, F., et al. (2024). An Advance Machine Learning (ML) Approaches for Anomaly Detection based on Network Traffic. Spectrum of Engineering Sciences, 2(3):502–527.

Abolmaali, S. M. A., Mohammadi, R., and Nassiri, M. (2024). IoT Malicious Traffic Classification and Detection Using Machine Learning Algorithms. In Development Engineering Conferences Center Articles Database, volume 1.

Ac¸ıkkar, M. and Tokgo¨z, S. (2025). Improving multi-class classification: scaled extensions of harmonic mean- based adaptive k-nearest neighbors. Applied Intelligence, 55(3):168.

AlSalehy, A. S. and Bailey, M. (2025). Improving Time Series Data Quality: Identifying Outliers and Handling Missing Values in a Multilocation Gas and Weather Dataset. Smart Cities, 8(3):82.

Alzaabi, F. R. and Mehmood, A. (2024). A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods. IEEE Access, 12:30907–30927.

Bala, B. and Behal, S. (2024). A Brief Survey of Data Preprocessing in Machine Learning and Deep Learning Techniques. In 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I- SMAC), pages 1755–1762. IEEE.

Diallo, R., Edalo, C., and Awe, O. O. (2024). Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score. In Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA, pages 283–312. Springer.

Eleftheriadis, C., Symeonidis, A., and Katsaros, P. (2024). Adversarial Robustness Improvement for Deep Neural Networks. Machine Vision and Applications, 35(3):35.

Garcıa, S., Parmisano, A., Erquiaga, M. J., Delgadillo, J., Hieskovsky´, M., Zonouzi, R., Cejka, T., Vojtech, J., DeKok, T., Sebestova, K., et al. (2020). Iot-23: A labelled dataset with malicious and benign IoT network traffic (Version 1.0.0)[Data set]. Zenodo.

Imani, M., Beikmohammadi, A., and Arabnia, H. R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. Technologies, 13(3):88.

Inuwa, M. M. & Das, R. (2024). A comparative analysis of various machine learning methods for anomaly detection in cyber-attacks on IoT networks. Internet of Things, 26:101162.

Jabardi, M. (2025). Support Vector Machines: Theory, Algorithms, and Applications. Infocommunications Journal, 17(1).

Kamal, H. and Mashaly, M. (2025). Robust Intrusion Detection System Using an Improved Hybrid Deep Learning Model for Binary and Multi-Class Classification in IoT Networks.

Technologies (2227-7080), 13(3).

Khan, S. A. (2024). Privacy-Preserving Deep Learning Framework for IoT Malware Detection. PhD thesis, Old Dominion University.

Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M., and Zschech, P. (2025). Challenging the Performance-Interpretability Trade-Off: An Evaluation of Interpretable Machine Learning Models. Business & Information Systems Engineerin. (68) pages 1–25.

Li, J. (2024). Area under the ROC Curve has the most consistent evaluation for binary classification. PLOS ONE, 19(12):e0316019.

Malalha, S. A. K., Burhanuddin, M., and Yunos, N. B. M. (2024). Unveiling the Tapestry of Machine Learning: A Comparative Analysis of Support Vector Machines, Random

Forests, and Neural Networks in Diverse Applications. Tuijin Jishu/Journal of Propulsion Technology, 45(3):2024.

Mohale, V. Z. and Obagbuwa, I. C. (2025). Evaluating machine learning-based intrusion detection systems with explainable ai: enhancing transparency and interpretability. Frontiers in Computer Science, 7:1520741.

Ness, S., Eswarakrishnan, V., Sridharan, H., Shinde, V., Janapareddy, N. V. P., and Dhanawat, V. (2025). Anomaly Detection in Network Traffic Using Advanced Machine Learning Techniques. IEEE Access. vol 13. pp 16133-16149.

Neto, E., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., and Ghorbani, A. A. (2023). Ciciot2023: CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. Sensors, 23(13):5941.

Rimal, Y., Sharma, N., and Alsadoon, A. (2024). The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. Multimedia Tools and Applications, 83(30):74349–74364.

Rimal, Y., Sharma, N., Paudel, S., Alsadoon, A., Koirala, M. P., and Gill, S. (2025). Comparative analysis of heart disease prediction using logistic regression, svm, knn, and random forest with cross-validation for improved accuracy. Scientific Reports, 15(1):13444.

Shafiq, M., Gu, Z., Cheikhrouhou, O., Alhakami, W., and Hamam, H. (2022). The Rise of "Internet of Things": Review and Open Research Issues Related to Detection and Prevention of IoT-Based Security Attacks. Wireless Communications and Mobile Computing, 2022(1):8669348.

Sujon, K. M., Hassan, R. B., Towshi, Z. T., Othman, M. A., Samad, M. A., and Choi, K. (2024). When to Use Standardization and Normalization: Empirical Evidence From Machine Learning Models and XAI. IEEE Access. vol. 12, pp. 135300-135314

Wiggerthale, J., & Reich, C. (2024). Explainable Machine Learning in Critical Decision Systems: Ensuring Safe Application and Correctness. *AI*, *5*(4), 2864-2896. https://doi.org/10.3390/ai5040138.

Zhang, HJ., Chen, CC., Ran, P. *et al.* A multi-dimensional hierarchical evaluation system for data quality in trustworthy AI. *J Big Data* **11**, 136 (2024). https://doi.org/10.1186/s40537-024-00999-2.

Zhang, W. & Lazaro, J. P. (2024). A Survey on Network Security Traffic Analysis and Anomaly Detection Techniques. International Journal of Emerging Technologies and Advanced Applications, 1(4):8–16.

Zhu, J., Pu, S., He, J., Su, D., Cai, W., Xu, X., and Liu, H. (2024). Processing imbalanced medical data at the data level with assisted-reproduction data as an example. BioData Min, 17(1):29.