

UVIF-IoT: Calibrated and Feature-Efficient Intrusion Detection with Forensic Triage for IoT Edge Security

Amal Ajayan^{*,†,‡}, Habib Hamam^{‡,§}, *Senior Member, IEEE*, and Mohsen Guizani^{*}, *Fellow, IEEE*

^{*}Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

[†]Amrita School of Physical Sciences Coimbatore, Amrita Vishwa Vidyapeetham, Tamil Nadu 641112, India

[‡]Faculty of Engineering, University of Moncton, Moncton, NB E1A 3E9, Canada

[§]School of Electrical Engineering, University of Johannesburg, Johannesburg 2006, South Africa

Emails: {amal.ajayan, mohsen.guizani}@mbzuai.ac.ae, Habib.Hamam@umoncton.ca

Abstract—Intrusion detection systems (IDSs) for Internet of Things (IoT) edge deployments must provide accurate detection while remaining compact, calibrated, robust, and useful for decision making. This paper proposes a UVIF-IoT, a feature-efficient and cyberforensics-supportive IDS pipeline for IoT edge security. UVIF-IoT combines Random-Forest-based feature ranking, compact neural classification, temperature-scaled confidence calibration, feature-space adversarial stress testing, and structured event triage. Evaluated on the N-BaIoT dataset with Mirai and Gafgyt botnet traffic, UVIF-IoT reduces the input dimensionality from 115 to 30 features (73.9% reduction). Compared with a full-feature multilayer perceptron (MLP), the compact model reduces trainable parameters from 23,234 to 4,130 and serialized model size from 96.39 KB to 20.10 KB, while achieving an F1-score of 0.999651 and an expected calibration error (ECE) of 0.000068 after temperature scaling. Under Projected Gradient Descent (PGD) feature-space perturbation at $\epsilon=0.02$, UVIF-IoT achieves an F1-score of 0.999453, versus 0.999306 for the full-feature MLP. UVIF-IoT also generates an event ledger for all 212,278 test samples, including source provenance, device identifier, attack type, calibrated attack probability, entropy, and a priority score for cyberforensics-based triage. The results show that compact IoT-edge intrusion detection can be combined with calibrated uncertainty, feature-space robustness evaluation, and analyst-facing forensic triage while maintaining a high detection performance.

Index Terms—Internet of Things security, intrusion detection systems, edge computing, N-BaIoT, feature selection, temperature scaling, calibration, uncertainty estimation, feature-space robustness, cyberforensics, botnet detection.

I. INTRODUCTION

The rapid proliferation of Internet of Things (IoT) devices has fundamentally transformed the threat landscape of modern networks. Billions of resource-constrained devices ranging from smart-home appliances and security cameras to industrial sensors are now operating across highly interconnected IoT environments, creating an expansive and heterogeneous attack surface [1]. Among the most consequential threats are IoT botnets, particularly the Mirai and Gafgyt/Bashlite families, which exploit weak configurations and the limited security capabilities of resource-constrained devices to conduct large-scale Distributed Denial-of-Service (DDoS) attacks [2].

Deploying effective Intrusion Detection Systems (IDSs) at the IoT edge is essential for detecting malicious activity while accommodating the computational limitations of resource-constrained devices [3]. However, IoT-edge IDS design faces a fundamental challenge: models that deliver high detection accuracy on centralized hardware may exceed the memory, computational, and storage budgets of constrained edge devices. Consequently, practical IDS design must jointly consider detection performance, model compactness, inference latency, and analyst-interpretable outputs [4].

Despite recent advances, three practically important dimensions continue to receive comparatively limited attention in IoT IDS research. First, the *calibration* of model confidence is evaluated less frequently than conventional detection metrics. Deep neural networks can produce poorly calibrated confidence scores, meaning that their predicted probabilities do not consistently correspond to empirical correctness likelihoods [5]. This limitation reduces the reliability of confidence-based incident prioritization. Second, the *forensic utility* of IDS outputs receives comparatively limited attention. Digital forensic readiness in IoT environments remains challenging because conventional centralized and static forensic procedures are poorly suited to distributed and heterogeneous IoT infrastructures [6]. Accordingly, IDSs that generate structured triage artifacts can help connect automated detection with analyst-led evidence collection and post-incident investigation. Third, feature-space robustness remains a critical concern because Deep Learning (DL)-based IDS models are susceptible to gradient-based adversarial perturbations, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) [7].

This paper proposes UVIF-IoT, a calibrated and feature-efficient IDS pipeline with cyberforensics-supportive triage for IoT edge security. UVIF-IoT jointly targets detection performance, calibrated uncertainty estimation, feature-efficient edge inference, feature-space robustness assessment, and triage output generation that supports cyberforensics-oriented analyst review.

A. Key Contributions

The key contributions of this work are as follows:

- 1) Feature-efficient compact detector: Random Forest (RF)-ranked feature selection reduces the N-BaIoT input from 115 to 30 features (73.9% reduction), followed by a compact MLP. UVIF-IoT achieves an F1-score of 0.999651 while reducing trainable parameters from 23,234 to 4,130 and serialized model size from 96.39 KB to 20.10 KB.
- 2) Post-hoc calibration via temperature scaling: Temperature scaling reduces ECE from 0.000218 to 0.000068 without altering the F1-score, improving the confidence calibration for analyst-facing use.
- 3) Adversarial feature-space stress test: UVIF-IoT is evaluated under FGSM and PGD feature-space perturbations and remains comparable to the full-feature MLP, achieving a marginally higher F1-score at PGD $\epsilon=0.02$ (0.999453 vs. 0.999306), showing that feature reduction does not degrade feature-space stability.
- 4) Cyberforensics-supportive triage: A structured event ledger is generated for all test samples, incorporating source-file provenance, device identifier, attack type, calibrated probability, predictive entropy, and a transparent composite priority score for triage.

The remainder of this paper is organized as follows. Section II reviews related literature. Section III details the UVIF-IoT methodology. Section IV describes the experimental setup. Section V presents and discusses the results. Section VI concludes the paper with future directions.

II. LITERATURE REVIEW

A. IoT Intrusion Detection

Machine learning (ML) and DL have been extensively investigated for IoT intrusion detection, with recent research increasingly considering not only predictive performance but also scalability and practical deployment. Hossain [8] compared one-dimensional convolutional neural networks (1D-CNNs), Long Short-Term Memory (LSTM) networks, recurrent neural networks (RNNs), and multilayer perceptrons (MLPs) on the CIC IoT-DIAD 2024 dataset, identifying the 1D-CNN as the strongest model for binary and multiclass IoT-traffic classification. For resource-limited edge environments, Ieropoulos et al. [3] developed a collaborative anomaly-detection architecture that separates centralized model training from lightweight distributed inference on IoT devices. Their system enables edge devices to exchange predictions and identify compromised nodes while respecting strict memory and computational constraints. Dinler [9] proposed task-scaled spiking neural networks for edge-compatible intrusion detection, reducing the Edge-IIoTset representation from 97 to 35 features and evaluating compact architectures across binary, six-class, and fifteen-class settings.

On N-BaIoT specifically, Meidan et al. [10] introduced the dataset and demonstrated network-based detection of Mirai and Bashlite botnet activity using deep autoencoders. More

recently, Umair et al. [4] evaluated knowledge distillation on N-BaIoT and CIC-IDS2023, showing that a compact student model can retain high detection performance while reducing parameter count, model size, and inference latency; their integration of SHAP also provides analyst-interpretable feature attributions. These studies illustrate the progression of IoT IDS research from benchmark-oriented detection toward collaborative, compact, edge-deployable, and explainable architectures.

B. Feature Selection and Lightweight IDS

High-dimensional traffic representations can impose computational, memory, and storage overheads that are prohibitive for resource-constrained edge deployment. Amr et al. [11] proposed a sequential submodular feature-sample selection framework that jointly reduces input dimensionality and training-set size. Evaluations on RT-IoT2022, Edge-IIoTset, and CICIoT2023 reduced the representation to 20 features and the training set to as little as 5% of its original size, providing training speedups of up to $70\times$ and reductions of up to 96% in peak training memory. Dinler [9] similarly reduced the Edge-IIoTset representation from 97 to 35 features and developed task-scaled spiking neural networks with memory footprints ranging from 36.5 KB for binary detection to 200.6 KB for fifteen-class classification.

Physical-device evaluations provide further evidence of deployability. Musthafa et al. [12] combined feature selection, pruning, and quantization to implement an ensemble IDS on resource-constrained Raspberry Pi hardware while maintaining 97.3% detection accuracy. Javed et al. [13] implemented a lightweight ML-based IDS for smart-home IoT devices, reporting 99.50% accuracy and an inference latency of 3.51 ms on embedded hardware. In a federated setting, Albanbay et al. [14] evaluated multiple learning architectures on Raspberry Pi 5 clients and found that a CNN provided the most favorable accuracy-latency trade-off, achieving approximately 98% accuracy. Collectively, these studies demonstrate the complementary roles of feature reduction, architecture compression, and hardware-level validation in establishing the practical suitability of IDS models for IoT-edge deployment.

C. Calibration, Feature-Space Robustness, and IoT Forensics

Beyond conventional classification performance, trustworthy IoT intrusion detection requires reliable confidence estimates, robustness to adversarial feature manipulation, and outputs that support post-incident investigation. Pal et al. [15] recently investigated calibrated uncertainty estimation specifically for deep IoT attack detection, demonstrating the importance of aligning predictive confidence with empirical reliability. More broadly, Dong et al. [5] surveyed confidence calibration under class imbalance, a particularly relevant setting for IDS datasets in which attack classes are commonly underrepresented.

Feature-space robustness constitutes a second concern. Yao et al. [16] modeled realistic adversarial traffic against DL-based IDSs in IIoT environments, showing that operationally plausible feature perturbations can substantially undermine

detection performance. Maseno et al. [7] further evaluated FGSM and PGD attacks against recurrent IDS models and used explainable AI to characterize changes in model behavior under adversarial evasion. These findings indicate that clean-data accuracy alone is insufficient for assessing the dependability of an IoT IDS.

Digital-forensic readiness constitutes a third, complementary requirement. KEBande and Awad [17] surveyed security and digital forensics across IIoT ecosystems, identifying persistent challenges associated with heterogeneous evidence sources and dynamic operating environments. Kuku et al. [6] subsequently proposed an organizational readiness and resilience model emphasizing incident handling, evidence acquisition, and systematic log collection. Collectively, these studies motivate IDS outputs that combine calibrated confidence, uncertainty indicators, provenance information, and structured event metadata to support both real-time triage and subsequent forensic investigation.

D. Summary of Research Gaps

The reviewed literature indicates four gaps that UVIF-IoT is designed to address.

Gap 1: Co-design of detection and edge efficiency: Although recent studies increasingly address edge efficiency, it is still commonly evaluated through isolated dimensions such as feature reduction, model compression, or hardware latency. Fewer studies jointly report detection performance, feature count, parameter count, serialized model size, and inference latency within one pipeline. UVIF-IoT integrates feature selection with a compact MLP and reports these complementary detection and efficiency measures.

Gap 2: Limited confidence calibration analysis. Conventional IoT IDS evaluations predominantly emphasize detection metrics such as accuracy, precision, recall, and F1-score, whereas the reliability of the associated confidence estimates is less frequently examined. UVIF-IoT addresses this by applying validation-based temperature scaling and reporting ECE before and after calibration; NLL and Brier score are also computed for completeness.

Gap 3: Robustness under compact models: The effect of feature reduction on neural IDS robustness under matched feature-space perturbations remains insufficiently characterized relative to corresponding full-feature models. UVIF-IoT directly compares compact and full-feature neural models under matched FGSM and PGD feature-space perturbation budgets.

Gap 4: Structured forensic triage output: The reviewed studies primarily report aggregate detection or deployment metrics, while structured per-sample event ledgers combining provenance, calibrated attack probability, entropy, and transparent priority scores remain comparatively underexplored.

III. METHODOLOGY

A. Framework Overview

The UVIF-IoT pipeline is summarized in Fig. 1. It begins with N-BaIoT data collection and preprocessing (missing-

value handling, median imputation, standardization, and stratified splitting), followed by RF-based feature ranking to select the top 30 features. A compact MLP is then trained, and temperature scaling is fitted using the validation set to obtain calibrated probabilities. The calibrated detector is evaluated under clean inference, CPU-latency measurement, and feature-space adversarial stress tests. Finally, UVIF-IoT produces a cyberforensics-supportive event ledger containing provenance, device identifier, attack type, calibrated probability, predictive entropy, and a triage priority score.

B. Problem Formulation

Let $\mathbf{x}_i \in \mathbb{R}^d$ denote an IoT traffic feature vector and $y_i \in \{0, 1\}$ its binary label (0 = benign, 1 = attack). In N-BaIoT, $d = 115$. UVIF-IoT is designed to jointly support accurate detection, calibrated confidence estimation, feature-efficient inference, feature-space robustness assessment, and triage supportive of cyberforensics.

C. Data Handling and Provenance

The dataset loader recursively discovers N-BaIoT CSV files and assigns binary labels from file names. For each sample, non-feature provenance fields (source file, device identifier, attack type) are preserved for triage outputs but excluded from model training. The data are split using stratified sampling.

To reduce leakage risk, all preprocessing statistics are estimated only from the training split. Median imputation and standardization are fitted on training data and then applied to validation and test data. RF-based feature ranking is also fitted only on the training data, and the selected feature set is fixed for validation and test evaluation. Temperature scaling is fitted only on the validation set, while the test set is used only for final evaluation.

D. Baseline Models

UVIF-IoT is compared against Logistic Regression (LR) (lightweight linear baseline), RF (strong tabular baseline and feature-ranking mechanism), and a full-feature MLP (all 115 N-BaIoT features, two hidden layers with ReLU activations and dropout). Neural models are trained using cross-entropy loss with early stopping based on validation F1-score.

E. Feature-Efficient UVIF-IoT

Let $\mathcal{F} = \{f_1, \dots, f_d\}$ denote the full feature set, where d is the original number of input features, and let $I(f_j)$ denote the Random Forest importance score of feature f_j . The top- k features are retained as

$$\mathcal{F}_k = \{f_j \in \mathcal{F} \mid I(f_j) \text{ is among the top } k \text{ scores}\}. \quad (1)$$

where $\mathcal{F}_k$ contains the k features with the largest RF importance scores. The feature reduction ratio is defined as

$$R_f = 1 - \frac{k}{d}, \quad (2)$$

where R_f measures the fraction of removed input features. In this work, $d = 115$ and $k = 30$, giving $R_f = 73.9\%$. The

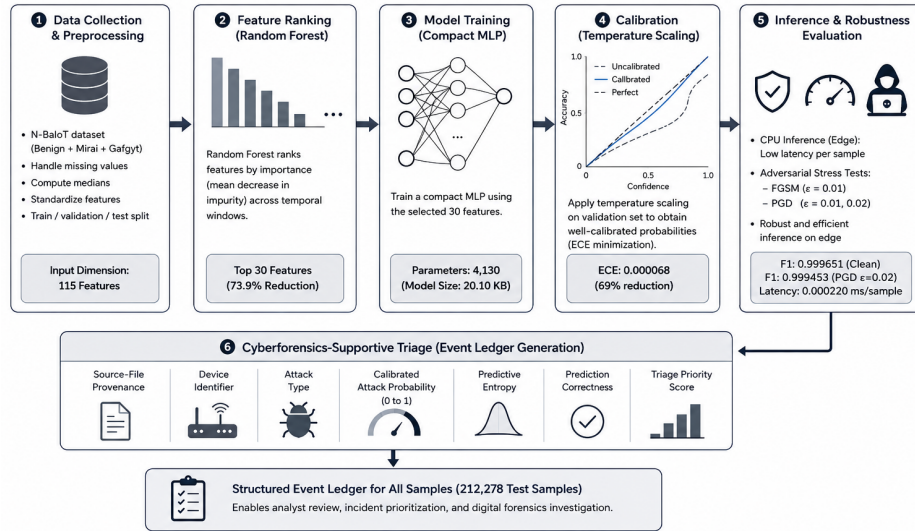


Fig. 1. Framework overview of the proposed UVIF-IoT pipeline.

selected features are used to train a compact MLP, reducing both input dimensionality and trainable parameters relative to the full-feature neural baseline.

F. Calibration and Uncertainty Estimation

Temperature scaling is fitted using the validation set. Given logits $\mathbf{z}$ and scalar temperature $T > 0$, calibrated probabilities are computed as

$$\mathbf{p} = \text{softmax}\left(\frac{\mathbf{z}}{T}\right), \quad (3)$$

where $\mathbf{p}$ is the calibrated class-probability vector and T is optimized by minimizing validation negative log-likelihood. Calibration is primarily reported using expected calibration error (ECE), while NLL and Brier score are also computed for completeness.

For each event, predictive entropy is computed as

$$H(\mathbf{p}) = -\sum_{c=1}^C p_c \log(p_c), \quad (4)$$

where C is the number of classes and p_c is the predicted probability of class c . The entropy is normalized by $\log C$ to obtain $H_{\text{norm}} \in [0, 1]$.

G. Cyberforensics-Supportive Triage

Each event ledger entry contains the event identifier, source file, device identifier, attack type, true and predicted labels, attack probability, confidence, entropy, prediction correctness, and a composite priority score:

$$S = 0.50 \cdot P(\text{attack}) + 0.30 \cdot H_{\text{norm}} + 0.20 \cdot (1 - \max_c p_c). \quad (5)$$

Here, S is the triage priority score, $P(\text{attack})$ is the calibrated attack probability, H_{norm} is the normalized predictive entropy, and $\max_c p_c$ is the prediction confidence. The generated outputs include top-priority events, high-uncertainty events, false

negatives, false positives, and device/attack-level summaries. These outputs are intended for cyberforensics-supportive triage and analyst review, not as a replacement for a complete legal forensic investigation.

H. Robustness Evaluation

FGSM and PGD are applied to standardized feature vectors. For perturbation budget ϵ , FGSM constructs

$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x}), y)), \quad (6)$$

where $\mathbf{x}_{adv}$ is the perturbed input, $\mathbf{x}$ is the original standardized feature vector, ϵ controls the perturbation magnitude, $\mathcal{L}$ is the classification loss, and y is the true label. PGD applies this update iteratively with projection back into the ϵ -bounded neighborhood of the original input. These experiments are controlled feature-space stability tests and are not interpreted as physical packet-level attacks.

IV. EXPERIMENTAL SETUP

A. Dataset and Evaluation Protocol

The experiments use the N-BaIoT dataset [10] for binary IoT intrusion detection. Binary detection is treated as the first-stage edge IDS task, where the immediate objective is to distinguish benign from malicious traffic before deeper attack-family analysis. The final run uses 1,061,384 balanced records (530,692 benign and 530,692 attack), split into 742,968 training, 106,138 validation, and 212,278 test samples via stratified sampling. Median imputation and standardization are fitted on the training set only.

B. Compared Models and Metrics

The compared models are LR, RF, a full-feature MLP using all 115 features, and UVIF-IoT using the top 30 RF-ranked features with a compact MLP. Detection is evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and

PR-AUC. Calibration is evaluated using ECE as the main reported metric, while NLL and Brier score are also computed. Robustness is evaluated using FGSM and PGD feature-space perturbations. Edge suitability is assessed using feature count, parameter count, serialized model size, and inference latency. Triage behavior is assessed through the generated event ledger.

V. RESULTS AND DISCUSSION

A. Clean Detection and Calibration

Table I reports clean detection and calibration results. RF achieves the highest clean F1-score of 0.999958. The full-feature MLP obtains 0.999637, while UVIF-IoT obtains 0.999651 using only 30 features. Temperature scaling reduces UVIF-IoT's ECE from 0.000218 to 0.000068 while leaving the test F1-score unchanged, indicating improved confidence calibration at no cost in detection performance.

TABLE I
CLEAN DETECTION AND CALIBRATION RESULTS ON N-BaIoT.

Model	F1	ROC-AUC	PR-AUC	ECE
LR	0.999143	0.999897	0.999928	0.000298
RF	0.999958	0.999995	0.999995	0.000058
MLP	0.999637	0.999981	0.999966	0.000224
UVIF-IoT	0.999651	0.999982	0.999983	0.000068

B. Robustness under Feature-Space Perturbations

Table II reports robustness under FGSM and PGD. Both neural models remain stable across the evaluated budgets. At PGD $\epsilon = 0.02$, UVIF-IoT achieves F1 0.999453 vs. 0.999306 for the full-feature MLP.

TABLE II
ROBUSTNESS UNDER FEATURE-SPACE PERTURBATIONS.

Model	Clean subset	FGSM@0.01	PGD@0.01	PGD@0.02
MLP	0.999619	0.999551	0.999551	0.999306
UVIF-IoT	0.999648	0.999541	0.999541	0.999453

Fig. 2 shows the PGD robustness curve. The compact UVIF-IoT model maintains comparable stability under the tested feature-space perturbations despite using substantially fewer features.

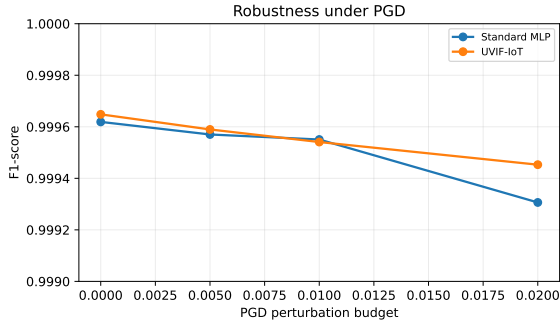


Fig. 2. PGD robustness comparison between full-feature MLP and UVIF-IoT under standardized feature-space perturbations.

C. Edge Efficiency

Table III compares edge-relevant efficiency metrics. UVIF-IoT reduces features by 73.9%, parameters from 23,234 to 4,130, serialized size from 96.39 KB to 20.10 KB, and CPU latency from 0.000401 to 0.000220 ms/sample.

TABLE III
EDGE-EFFICIENCY COMPARISON BETWEEN FULL-FEATURE MLP AND UVIF-IoT.

Model	Feat.	Red.(%)	Params	Size(KB)	CPU Lat.(ms)
MLP	115	0.0	23,234	96.39	0.000401
UVIF-IoT	30	73.9	4,130	20.10	0.000220

D. Cyberforensics-Supportive Triage

UVIF-IoT generates an event ledger for all 212,278 test samples. Table IV summarizes the triage outputs. The generated files support analyst review by identifying suspicious or uncertain events and summarizing model behavior across devices and attack types. The priority score is used as a transparent triage heuristic and is not claimed to be a user-validated analyst utility model.

TABLE IV
CYBERFORENSICS-SUPPORTIVE TRIAGE SUMMARY.

Item	Value
Test events	212,278
False negatives	32
False positives	42
Mean normalized entropy	0.003352
Priority score	Risk + uncertainty

E. Overall Discussion

UVIF-IoT achieves a practical trade-off: RF delivers the highest clean detection score, but UVIF-IoT obtains comparable neural detection with substantially fewer features and a much smaller model. Binary detection on N-BaIoT is highly separable, and all compared models operate near the performance ceiling; the contribution therefore lies not in raw detection gains but in reaching that ceiling with a fraction of the features, parameters, and model size, together with calibrated confidence and triage support. Compared with the full-feature MLP, UVIF-IoT reduces the feature count, parameter count, model size, and latency while maintaining competitive detection, calibration, and robustness. The main contribution is therefore a compact and reproducible IDS pipeline that integrates feature-efficient inference, calibrated uncertainty, feature-space robustness testing, and triage outputs that support cyberforensics.

Two limitations should be noted: the split is stratified by label rather than held out by device or attack family, and the balanced setting does not reflect the skewed class priors of operational IoT deployments, so the calibration results should be interpreted under the balanced evaluation protocol.

VI. CONCLUSION

This study presented UVIF-IoT, a feature-efficient, calibrated, and cyberforensics-supportive intrusion detection pipeline for IoT-edge security. The framework combines RF-based feature ranking, compact neural classification, temperature-scaled confidence calibration, feature-space adversarial stress testing, and structured event triage. Unlike IDS pipelines that focus primarily on clean classification accuracy, UVIF-IoT is designed to address practical deployment requirements, including reduced feature cost, compact inference, calibrated confidence estimation, robustness assessment, and analyst-facing review.

The results show that UVIF-IoT can preserve high detection performance while reducing the input feature set and neural model footprint. The calibrated confidence scores and structured event ledger further support cyberforensics-oriented triage by identifying suspicious, uncertain, and misclassified events with preserved source, device, and attack-type provenance. These properties make UVIF-IoT suitable for lightweight IoT-edge monitoring scenarios where detection accuracy alone is insufficient.

A. Future Work

Future work will evaluate UVIF-IoT using held-out device-wise and attack-wise partitions to assess generalization to previously unseen devices and attack categories. Evaluation under imbalanced and deployment-prior class distributions will provide a more realistic characterization of operational performance. Robustness analysis will also be extended to traffic-space adversarial scenarios that preserve protocol semantics and network functionality. In addition, adaptive calibration methods will be investigated under device, attack, and class-prior shifts. The forensic utility of the framework will be examined through analyst-centered validation of the proposed priority score and the integration of mechanisms for maintaining the integrity and provenance of generated triage artifacts. Finally, federated deployment across distributed IoT edge nodes will be explored, with particular attention to client heterogeneity, communication overhead, aggregation stability, and data privacy.

The source code, experimental notebook, configuration files, generated tables, figures, logs, and triage artifacts used in this study are publicly available in the project repository [18].

REFERENCES

- [1] B. Gușă, A. A. Anton, C. S. Stângăciu, D. Stănescu, L. I. Găină, and M. V. Micea, "Securing IoT edge: A survey on lightweight cryptography, anonymous routing and communication protocol enhancements," *International Journal of Information Security*, vol. 24, p. 149, 2025, <https://doi.org/10.1007/s10207-025-01071-7>.
- [2] U. Garg, P. Mishra, N. Gupta, and E. S. Pilli, "IoT botnets unveiled: Architectural analysis, threat vectors, and cutting-edge detection techniques," *Cluster Computing*, vol. 28, p. 945, 2025, <https://doi.org/10.1007/s10586-025-05633-1>.
- [3] V. Ieropoulos, E. Anthi, T. Spyridopoulos, P. Burnap, P. Carnelli, and A. Khan, "Federated detection at the edge: Collaborative anomaly detection for resource-limited IoT," *IEEE Internet of Things Journal*, vol. 13, no. 11, pp. 24 205–24 218, 2026, <https://doi.org/10.1109/JIOT.2026.3674644>.
- [4] M. Umair, M. S. Khan, W. A. Malwi, F. Asiri, I. Nafea, F. Saeed, and J. Ahmad, "Knowledge distillation for lightweight and explainable intrusion detection in resource-constrained consumer devices," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 4, pp. 12 157–12 165, 2025, <https://doi.org/10.1109/TCE.2025.3601183>.
- [5] J. Dong, Z. Jiang, D. Pan, Z. Chen, Q. Guan, H. Zhang, G. Gui, and W. Gui, "A survey on confidence calibration of deep learning-based classification models under class imbalance data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 9, pp. 15 664–15 684, 2025, <https://doi.org/10.1109/TNNLS.2025.3565159>.
- [6] O. Kuku, A. Chrysikos, and S. Salekzamanckhani, "Preparing IoT-enabled organisations for digital forensics: Model for readiness and resilience," *International Journal of Information Security*, vol. 24, p. 170, 2025, <https://doi.org/10.1007/s10207-025-01079-z>.
- [7] E. M. Maseno, Y. Sun, and Z. Wang, "Detecting adversarial evasion in deep learning intrusion detection systems using explainable AI," *International Journal of Information Security*, vol. 25, p. 125, 2026, <https://doi.org/10.1007/s10207-026-01299-x>.
- [8] M. A. Hossain, "Deep learning-based intrusion detection for IoT networks: A scalable and efficient approach," *EURASIP Journal on Information Security*, vol. 2025, p. 28, 2025, <https://doi.org/10.1186/s13635-025-00202-w>.
- [9] Ö. B. Dinler, "Task-scaled spiking neural networks with multi-stage feature selection for resource-constrained IoT intrusion detection," *Cluster Computing*, vol. 29, p. 198, 2026, <https://doi.org/10.1007/s10586-026-06064-2>.
- [10] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-BaIoT—network-based detection of IoT botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, 2018, <https://doi.org/10.1109/MPRV.2018.03367731>.
- [11] M. N. Amr, A. Elliethy, T. Mekawy, and A. Mahran, "Sequential submodular feature-sample selection for lightweight and update-efficient IoT intrusion detection," *Scientific Reports*, vol. 16, no. 1, p. 26315, 2026, <https://doi.org/10.1038/s41598-026-66614-x>.
- [12] M. B. Musthafa, S. Huda, T. T. Nguyen, Y. Kodera, and Y. Nogami, "Optimized ensemble deep learning for real-time intrusion detection on resource-constrained Raspberry Pi devices," *IEEE Access*, vol. 13, pp. 113 544–113 556, 2025, <https://doi.org/10.1109/ACCESS.2025.3584373>.
- [13] A. Javed, A. Ehtsham, M. Jawad, M. N. Awais, A.-u.-H. Qureshi, and H. Larjani, "Implementation of lightweight machine learning-based intrusion detection system on IoT devices of smart homes," *Future Internet*, vol. 16, no. 6, p. 200, 2024, <https://doi.org/10.3390/fi16060200>.
- [14] N. Albanbay, Y. Tursynbek, K. Graffi, R. Uskenbayeva, Z. Kalpeyeva, Z. Abilkaiyr, and Y. Ayapov, "Federated learning-based intrusion detection in IoT networks: Performance evaluation and data scaling study," *Journal of Sensor and Actuator Networks*, vol. 14, no. 4, p. 78, 2025, <https://doi.org/10.3390/jsan14040078>.
- [15] B. Pal, M. S. Islam, and A. W.-C. Liew, "Calibrated uncertainty estimation for trustworthy deep IoT attack detection," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 6, pp. 7571–7584, 2025, <https://doi.org/10.1109/TDSC.2025.3598350>.
- [16] W. Yao, H. Peng, Q. Li, and X. Shen, "Modeling realistic adversarial traffic against deep-learning-based intrusion detection system in industrial IoT," *IEEE Internet of Things Journal*, vol. 12, no. 15, pp. 29 540–29 554, 2025, <https://doi.org/10.1109/JIOT.2025.3568503>.
- [17] V. R. Kebande and A. I. Awad, "Industrial internet of things ecosystems security and digital forensics: Achievements, open challenges, and future directions," *ACM Computing Surveys*, vol. 56, no. 5, p. 131, 2024, <https://doi.org/10.1145/3635030>.
- [18] A. Ajayan, H. Hamam, and M. Guizani, "UVIF-IoT: Reproducibility package," <https://github.com/amalajayan13/UVIF-IoT>, 2026.