Proceedings of CONF-CDS 2026 Symposium: Data-Centric AI Security: Securing Models, Learning Agents, and Autonomous Systems
DOI: 10.54254/2755-2721/2026.CH35701

Random Forest Based Dual Attack Detection for Smart Internet of Things (IoT) Home Appliance

Jingtao Yan

School of Electronic Information, Xijing University, Xi'an, China
jingtaoyan0301@163.com

Abstract. With the rapid proliferation of IoT technology, smart home appliances face growing data security threats, especially model probing attacks and network-layer DDoS attacks. This paper proposes a machine learning based dual detection system. For model probing defense, we design a query-behavior detection framework extracting three core indicators (top-1 probability, margin, top-2 probability) and employ a Random Forest classifier. For IoT network attacks, we construct an end-to-end detection framework using multi-dimensional traffic features. Experiments on two public datasets show that the probing attack detector achieves 75.7% and 82.1% accuracy on DDoS_1 and DDoS_2, while the network attack detector achieves 76.53% accuracy and 0.76 F1-score. Feature importance analysis reveals key detection signals. We discuss performance bottlenecks and future optimization directions including data balancing, deeper feature engineering, deep learning models, and lightweight edge deployment.

Keywords: IoT security, Smart home appliances, Model probing attack, Random forest, Hybrid detection

1. Introduction

The Internet of Things (IoT) is transforming industrial and domestic environments. Smart home appliances, as key IoT carriers, enable remote control via cloud or gateways, forming complete data flows [1, 2]. However, the rapid expansion brings severe security challenges: over 70% of IoT devices are vulnerable [3], and in 2025 alone, security agencies detected 13.6 billion IoT attacks [4]. IoT devices often suffer from resource constraints and weak protection [5, 6], and manufacturers often prioritize production over security [7].

Two major threats are prominent. First, AI model assets face probing attacks, where adversaries repeatedly query the target model and reconstruct surrogate models from output predictions [8]. Second, IoT network infrastructures are threatened by DDoS attacks, while traditional intrusion detection systems struggle with scale and complexity [6, 9]. Given limited device resources, conventional heavyweight solutions are infeasible [5]. This paper proposes an intelligent data security detection system leveraging machine learning (Random Forest) for real-time identification of both threats [6, 9]. Hybrid approaches combining machine learning with nature-inspired algorithms and dual-stage pipelines have shown promise, and we build upon these ideas [10, 11].

DOI: 10.54254/2755-2721/2026.CH35701

Validation on public datasets includes feature importance analysis to support lightweight edge deployment [5, 6, 9].

2. Related technologies and analysis of smart home security threats

2.1. Foundation architecture of IoT smart home appliances

The fundamental architecture follows an end-to-end data flow from perception to application layers [1, 2]. Multiple attack surfaces exist (e.g., weak passwords, unencrypted streams, insecure APIs) [3]. Heterogeneity complicates unified protection [2], and machine learning models have been effective on datasets like CICIoT2023 [12].

2.2. Core detection technical principles

Random Forest is an ensemble method constructing multiple decision trees and outputting majority vote [5]. It offers good interpretability, robustness, and low overhead, suitable for IoT environments [5, 6, 13].

Model Probing Attack: Attackers repeatedly query public APIs and collect returned probabilities (top-1, top-2) and margins to reconstruct surrogate models [8, 14].

DDoS Attack: Attackers use compromised devices to flood targets with malicious traffic, exhausting resources [2, 6, 9]. Lightweight IDS for MQTT-enabled devices have been proposed [15, 16].

DDoS Attack Principle: Distributed Denial of Service (DDoS) attacks aim to exhaust the computational or network resources of a target system by flooding it with a large volume of malicious traffic from multiple compromised devices [2]. In IoT ecosystems, vulnerable smart devices such as surveillance cameras and network routers are frequently compromised and recruited into botnets to initiate large-scale DDoS flooding attacks [6, 9]. These attacks can severely disrupt smart home services and compromise the stability of the entire IoT ecosystem [5]. Lightweight intrusion detection systems specifically designed for MQTT-enabled IoT devices have been proposed to address the unique challenges of smart home protocols [15, 16].

2.3. Analysis of two types of core security threats in smart homes

We use Accuracy, Precision, Recall, and F1-score, with attack samples as positive class. All evaluations on held-out test sets.

3. Design of dual-module security detection system for smart IoT home appliances

3.1. Overall system architecture

The system comprises two independent modules sharing underlying infrastructure: (1) application-layer model probing detection, analyzing query logs; (2) network-layer traffic attack detection, analyzing raw packets. Both use Random Forest with interpretable feature engineering and rule-based pre-filtering.

3.2. Design of model probing attack detection module

Three core features extracted from each query response [8, 14]:

DOI: 10.54254/2755-2721/2026.CH35701

(1) Top-1 probability (highest score);
(2) Margin (difference between highest and second-highest);
(3) Top-2 probability (second-highest).

Workflow: (1) Query log collection, (2) Feature extraction, (3) Random Forest inference, (4) Binary decision.

Parameters: n_estimators=80, max_depth=14, min_samples_leaf=2, class_weight="balanced" to handle imbalance.

3.3. Design of the IoT network traffic attack detection module

End-to-end framework directly extracts multi-dimensional features from raw packets after preprocessing (missing value imputation, scaling, one-hot encoding) [5, 6, 9]. Features include byte counts, connection flags, login counts, etc. Operates in real time for gateway/edge deployment.

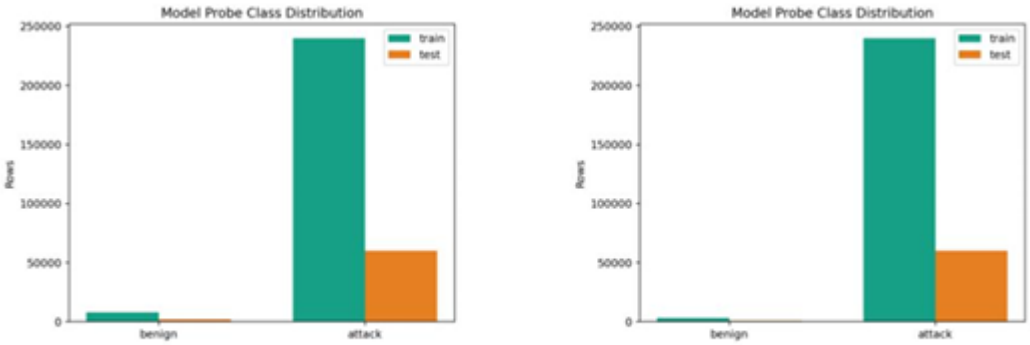
4. Experimental verification and result analysis

4.1. Experimental environment and dataset construction

Table 1. Model probing attack datasets (DDoS_1 and DDoS_2)

Dataset	Benign Samples (Test)	Attack Samples (Test)	Ratio (Attack:Benign)
DDoS_1	815	60000	73.6:1
DDoS_2	2000	60000	30:1

Two public datasets for model probing: DDoS_1 (test: 815 benign, 60,000 attack; ratio 73.6:1) and DDoS_2 (test: 2,000 benign, 60,000 attack; ratio 30:1). Both show severe class imbalance [8]. Training sets maintain similar proportions.



(a) DDOS_1 (b) DDOS_2

Figure 1. Sample distribution of training and test sets for model probing attack detection

Table 2. IoT network attack dataset

Dataset	Benign Samples	Attack Samples	Total
Training	68000	59000	125973
Test	10000	13000	22544

DOI: 10.54254/2755-2721/2026.CH35701

For IoT network attacks, a balanced dataset with training (68,000 benign, 59,000 attack) and test (10,000 benign, 13,000 attack) [5].

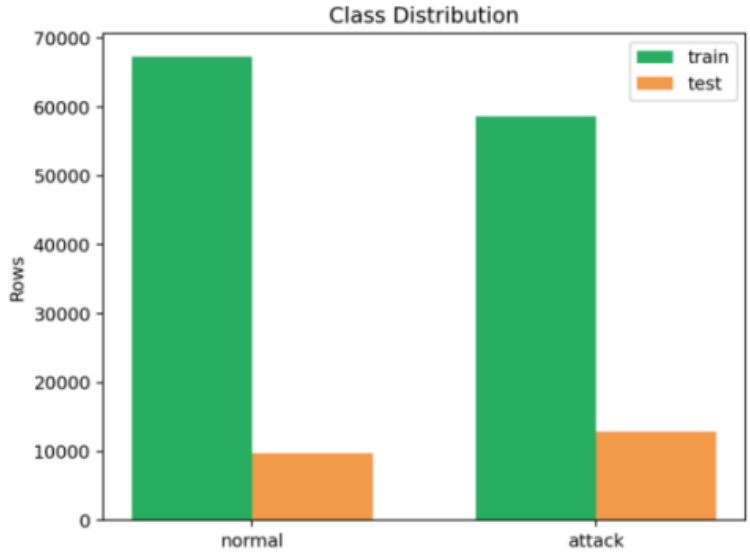


Figure 2. Training and evaluation of the IoT attack detection model

4.2. Analysis of experimental results for model-based attack detection

Table 3 summarizes all core detection metrics and feature importance rankings obtained on the DDoS_1 and DDoS_2 datasets for model probing attack detection.

Table 3. Performance and feature importance of model probing detection on two datasets

Index	DDos_1	DDoS_2
Overall Accuracy	75.7%	82.1%
Attack Recall	75.7%	82.3%
Benign Precision	0.04	0.12
Feature Importance Ranking	Top-1 probability (0.389) > margin (0.360) > top-2 probability (0.251)	Top2 probability (0.358) > top-1 probability (0.354) > margin (0.288)

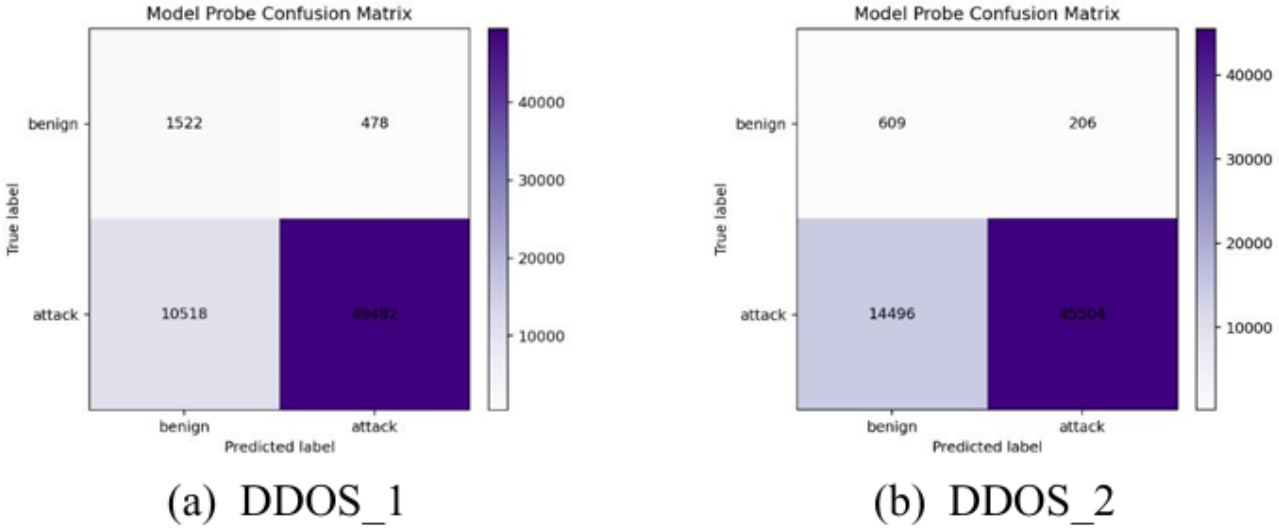
On DDoS_1: Accuracy 75.7%, Attack Recall 75.7%, Benign Precision 0.04 (96% of predicted benign are attacks). Feature importance: top-1 (0.389) > margin (0.360) > top-2 (0.251).

On DDoS_2: Accuracy 82.1%, Attack Recall 82.3%, Benign Precision 0.12. Feature importance: top-2 (0.358) ≈ top-1 (0.354) > margin (0.288).

Key Observations:

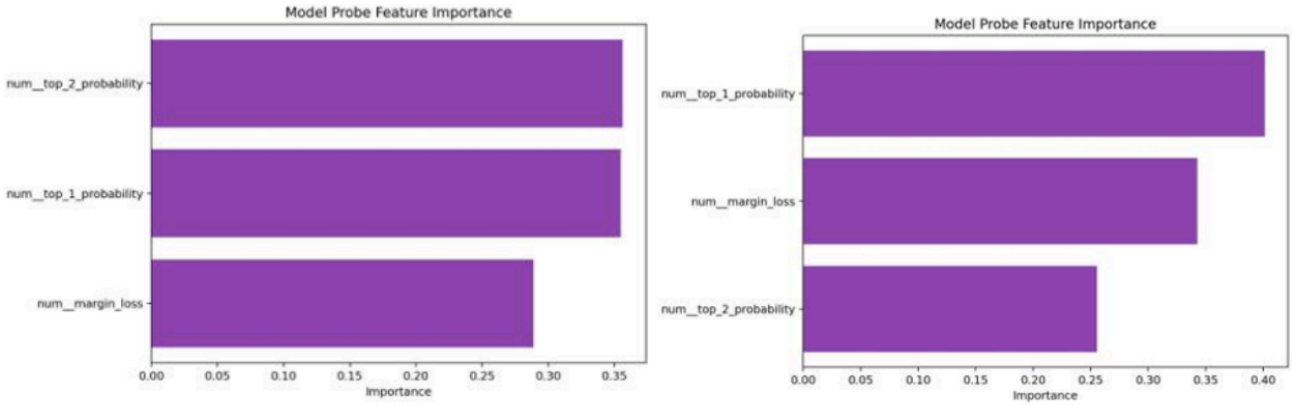
Severe class imbalance hurts benign identification; confusion matrices (Figure 3 omitted) show high false positives on benign samples. Feature importance varies slightly across datasets, but all three features collectively support detection [8, 14].

DOI: 10.54254/2755-2721/2026.CH35701



(a) DDOS_1 (b) DDOS_2

Figure 3. Confusion matrix for the model probing attack dataset



(a) DDOS_1 (b) DDOS_2

Figure 4. Feature importance plot for model probing attack detection

4.3. Analysis of experimental results for IoT ddos traffic attack detection

Overall Performance:Accuracy 76.53%, Attack Precision 96.77% (good false positive control), Attack Recall 60.80% (high miss rate), F1-score 0.76 [5, 6].

Confusion matrix (Figure 5 omitted):9,451 benign correctly identified, 260 false positives; 5,031 attacks missed (conservative flagging) [9].

Feature importance (Figure 6 omitted): source byte count (~0.17), destination byte count, and connection flag SF dominate; byte-level statistics are strongest signals [5, 6].

DOI: 10.54254/2755-2721/2026.CH35701

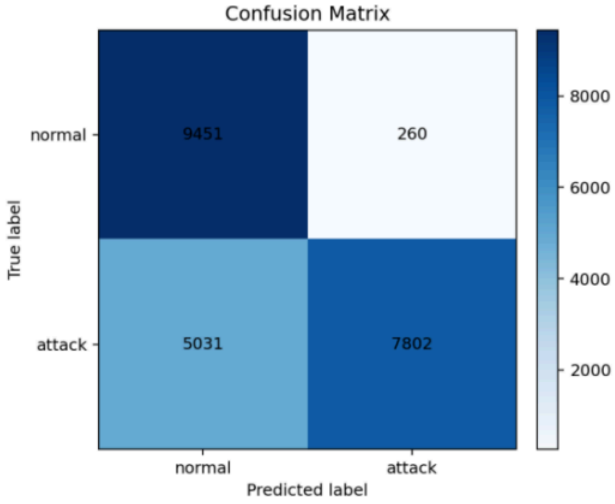


Figure 5. Confusion matrix for the IoT attack detection model

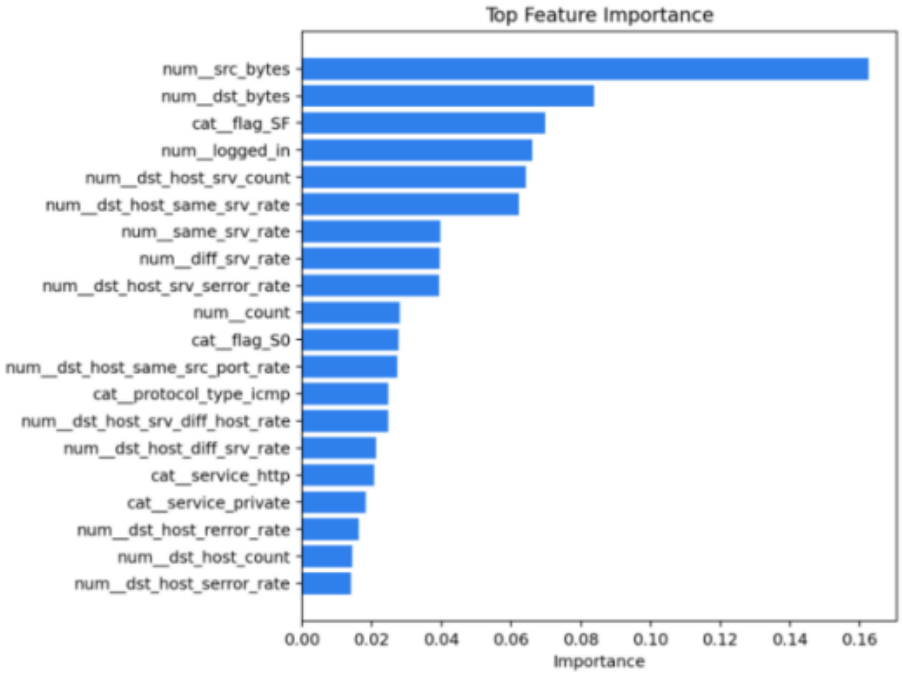


Figure 6. Feature importance plot for the IoT attack detection model

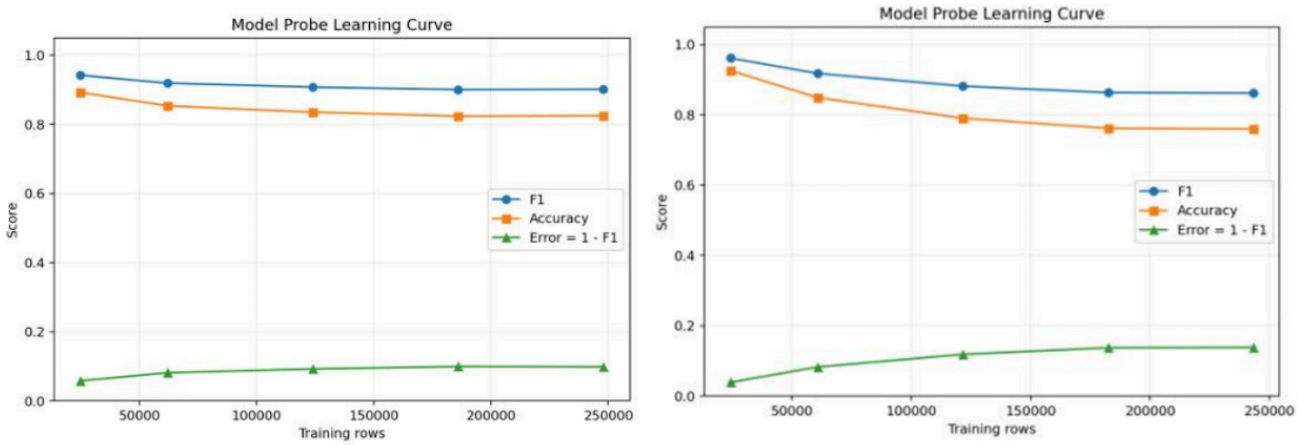
4.4. Analysis of the causes of existing performance defects in two types of detection tasks

Probing detection: Extremely low benign precision due to severe class imbalance (ratios 73.6:1 and 30:1), preventing learning of benign features [8]. Learning curves (Figure 7 omitted) show stable performance without overfitting for DDoS_1, but persistent fluctuations for DDoS_2, indicating insufficient benign characterization [14].

IoT Traffic Attack Detection: Low attack recall (60.8%) because the model is overly conservative, preferring to classify uncertain samples as benign [9]. Learning curve (Figure 8 omitted) shows saturation around 0.75-0.77 F1/accuracy even with increasing data, indicating bottleneck in feature discriminative power or task complexity [5, 6].

DOI: 10.54254/2755-2721/2026.CH35701

Common Issues Across Both Detection Tasks: Single Random Forest with fixed hyperparameters may not be optimal; current feature engineering insufficient; performance plateaus suggest need for deeper feature engineering or advanced architectures [9, 17]. Hybrid approaches with feature selection and ensemble methods could help [10, 18].



(a) DDOS_1 (b) DDOS_2

Figure 7. Learning curve for the model probing attack detection task

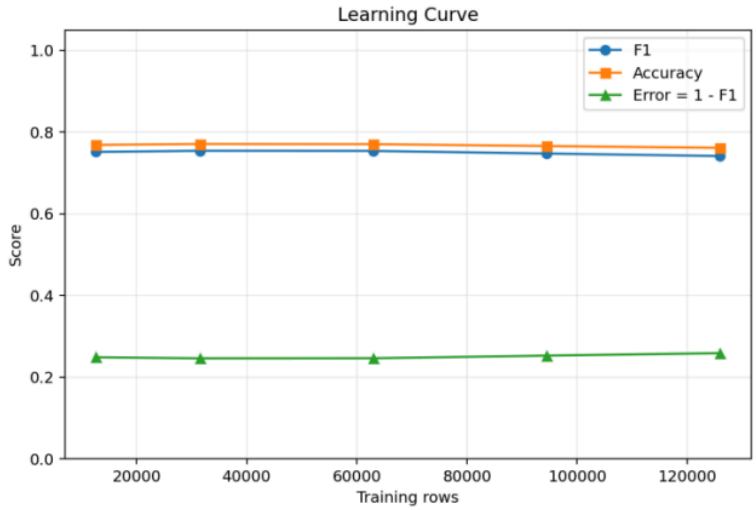


Figure 8. Learning curve for the IoT attack detection model

5. Conclusion

This paper proposes a machine learning based dual detection system for smart IoT home appliance security. For model probing attacks, we extract top-1 probability, margin, and top-2 probability, and use Random Forest to effectively identify malicious queries, validated across two datasets. For IoT network attacks, we build an end-to-end model using multi-dimensional traffic features, achieving 76.53% accuracy and 0.76 F1-score. Feature importance analysis reveals that model confidence changes are strongest for probing, while byte-level statistics are key for network attack detection.

Current system has limitations: probing detection suffers from high false positives due to data imbalance; network detection has high false negatives (low recall) and performance saturation.

DOI: 10.54254/2755-2721/2026.CH35701

Learning curves show that merely increasing data volume cannot further improve performance; the bottleneck lies in feature expressiveness and task complexity.

To address these, we propose four optimization directions:

(1) Data balancing: Use oversampling/undersampling or adjust decision thresholds to reduce false positives.

(2) Deeper feature engineering: Mine time-series features and sliding-window statistics to enhance pattern recognition.

(3) Deep learning models: Introduce RNN/LSTM to exploit sequential data, potentially reducing false negatives.

(4) Model lightweighting and deployment: Apply pruning/quantization for real-time edge deployment.

References

[1] *Research on intrusion detection methods for smart homes based on multi-layer neural networks* [Master's thesis]. Chongqing University of Posts and Telecommunications, 2024.

[2] *Design and implementation of security risk detection model for IoT rule linking based on graph representation learning* [Master's thesis]. 2023.

[3] *Intrusion detection in heterogeneous smart home IoT* [Master's thesis]. University of Oulu, 2026.

[4] *Research and application of intrusion detection methods for IoT* [Master's thesis]. 2024.

[5] Yan, S. L. *Research on IoT DDoS attack detection based on convolutional neural networks and federated learning* [Master's thesis]. Hubei University of Technology, 2024.

[6] *Research on lightweight intrusion detection methods for IoT* [Master's thesis]. 2025.

[7] *Creating a security conscious IoT ecosystem with innovative monitoring and classification solutions* [Doctoral thesis]. Óbuda University, 2025.

[8] Xiao, Y. X. *A study on learning-based model extraction attacks and defense methods* [Ph.D. thesis]. Hong Kong Polytechnic University, 2025.

[9] *Research on IoT DDoS attack detection methods based on deep learning* [Master's thesis]. 2024.

[10] A machine learning and nature inspired algorithm based hybrid approach for detection of IoT attacks. 2024 IEEE International Conference, Hooghly, India, Sep. 2024. (Date Added to IEEE Xplore: Jan. 2025)

[11] A dual-stage ML pipeline for IoT DDoS defense: Anomaly filtering and risk-prioritized detection. 2025 IEEE International Conference, Chennai, India, Aug. 2025. (Date Added to IEEE Xplore: Nov. 2025)

[12] Development of intrusion detection models for IoT networks utilizing CICIoT2023 dataset. 2023 IEEE International Conference, Bali, Indonesia, Dec. 2023. (Date Added to IEEE Xplore: Feb. 2024)

[13] Efficient and interpretable random forest-based attack detection for IoT network traffic. In *2025 30th International Conference on Information Technology (IT)*, IEEE, Subotica, Serbia, Sep. 2025. (Date Added to IEEE Xplore: Oct. 2025)

[14] Fei, T. (2025). *Research on model extraction attacks based on breaking entangled watermarks* [Master's thesis]. Xiangtan University.

[15] Lightweight intrusion detection system for MQTT-enabled IoT devices. In *2025 IEEE International Conference*, Lisbon, Portugal, Jul. 2025. (Date Added to IEEE Xplore: Sep. 2025)

[16] Machine learning classification for attack trace detection in smart home enabled IoT devices. In *2025 IEEE International Conference*, Coimbatore, India, Sep. 2025. (Date Added to IEEE Xplore: Dec. 2025)

[17] Next-generation IoT intrusion detection: Leveraging hybrid deep learning algorithms for enhanced security. In *2025 IEEE Conference*, Feb. 2025.

[18] Intrusion detection system traffic classification based on machine learning with correlation-based filtering and a genetic algorithm-inspired feature selection method for IoT networks. *Engineering, Technology & Applied Science Research, 15*(5), Oct. 2025.