STM JOURNALS

International Journal of
Wireless Security and Networks

ISSN: 2584-0703
Volume 4, Issue 1, 2026
January–June
DOI (Journal): 10.37591/IJWSN

https://journals.stmjournals.com/ijwsn

Research IJWSN

Assessing the Robustness of Machine Learning Models for Wireless Intrusion Detection Under Adversarial Traffic Perturbations

Siddhant Sukhatankar*

Abstract

As the Internet of Things (IoT) devices and wireless communication networks continue to grow rapidly, protecting systems from cyber threats has become increasingly important. Machine learning–based intrusion detection systems (IDS) have shown strong potential in detecting abnormal and malicious network activities, yet their effectiveness and resilience when facing adversarial attacks are still not sufficiently explored. This research evaluates Machine Learning (ML) models–XGBoost, random forest, and multi-layer perceptron (MLP)—in detecting attacks within wireless IoT networks when subjected to certain specific traffic feature perturbations. Using the CICIoT2023 and IoT intrusion datasets, we conducted binary classification experiments distinguishing benign and attack traffic. Perturbations simulating realistic adversarial manipulations were applied to numeric features at multiple levels (5%, 10%, and 25%). The results demonstrate that tree-based models, XGBoost, and random forest maintain high recall under perturbation, with less than 0.2% reduction at even the highest perturbation levels, whereas MLP performance is unstable on imbalanced data. Feature importance analysis reveals that timing and protocol-related features contribute significantly to model predictions. These findings highlight the robustness of ensemble tree methods in practical IoT intrusion detection scenarios and emphasize the need for perturbation-aware evaluations for ML-based IDS. The study contributes to safer deployment of wireless IDS and provides a methodological framework for assessing model resilience against adversarial feature modifications.

Keywords: Adversarial perturbations, IoT security, machine learning, random forest, wireless intrusion detection, XGBoost

INTRODUCTION

The modern Internet of Things (IoT) paradigm has transformed how devices communicate, offering substantial efficiency and convenience in homes, industries, and urban infrastructure [1]. However, the rapid adoption of IoT devices has increased the attack surface for malicious actors. Distributed Denial of Service (DDoS) attacks, Mirai botnets, and other intrusion strategies exploit vulnerabilities in network protocols and device firmware [2]. Traditional signature-based intrusion detection systems (IDS) methods struggle with novel attacks and high-volume traffic, motivating the use of machine learning models for anomaly and intrusion detection [3].

Machine learning models, including ensemble tree methods (e.g., random forest, XGBoost) and neural networks (e.g., multi-layer perceptron (MLP)), have been increasingly applied for IDS in IoT networks [4, 5]. These models can automatically learn patterns from high-dimensional traffic data, thereby enabling the detection of

***Author for Correspondence**
Siddhant Sukhatankar
E-mail: siddhantsukhatankar@gmail.com

[1]Software Development Engineer–Artificial Intelligence & Machine Learning (AI/ML), Demand Science Optimization Organization, Amazon, Arlington, Virginia, United States

Received Date: January 13, 2026
Accepted Date: January 16, 2026
Published Date: January 20, 2026

Citation: Siddhant Sukhatankar. Assessing The Robustness of Machine Learning Models for Wireless Intrusion Detection Under Adversarial Traffic Perturbations. International Journal of Wireless Security and Networks. 2026; 4(1): 29–34p.

previously unseen attacks. Despite their popularity, the robustness of these models under realistic adversarial conditions, such as traffic feature perturbations mimicking evasion attempts, has been overlooked [6].

Adversarial perturbations, typically studied in computer vision, involve small feature-level modifications designed to degrade the model's performance [7]. In network traffic, perturbations can arise from attackers manipulating packet timing, rates, or flag counts to evade detection. Evaluating ML models under such conditions is critical to ensure reliable deployment in real-world IoT networks.

This study investigated the robustness of widely used ML models in wireless intrusion detection when exposed to adversarial perturbations. This study aims to answer the question: "How robust are commonly used machine learning models for wireless intrusion detection when subjected to realistic adversarial perturbations in traffic features?"

MATERIALS AND METHODOLOGY

Datasets

Two datasets were utilized: CICIoT2023 [8] and IoT intrusion [9]. Both datasets contain flow-level traffic records with features representing the timing, protocol, statistical, and flag-based properties. The CICIoT2023 dataset includes 1,176,851 training samples, 1,176,851 validation samples, and 1,176,851 test samples, with 47 features, including flow duration, protocol type, Transmission Control Protocol (TCP)/ User Datagram Protocol (UDP) flags, and statistical measures (sum, mean, variance). IoT intrusion contains 1,048,575 samples with 47 features, including benign traffic and diverse attack types, as listed in Table 1.

The datasets were preprocessed to remove constant features (e.g., Telnet, IRC) and non-numeric columns. Labels were binarized into benign (0) and attack (1) categories. After preprocessing, the final feature count was 41 for both datasets.

Feature Engineering and Scaling

- *Timing features*: flow_duration, duration, inter-arrival time (IAT), rate, Srate, Drate
- *Flag features*: fin_flag_number, syn_flag_number, rst_flag_number, psh_flag_number, ack_flag_number, ece_flag_number, and cwr_flag_number
- *Statistical features*: Tot_sum, min, max, AVG, Std, Tot_size, magnitude, radius, covariance, variance, weight
- *Protocol features*: Transmission Control Protocol (TCP), User Datagram Protocol (UDP), Hypertext Transfer Protocol (HTTP), Hypertext Transfer Protocol Secure (HTTPS), Domain Name System (DNS), Simple Mail Transfer Protocol (SMTP), Secure Shell (SSH), Dynamic Host Configuration Protocol (DHCP), Address Resolution Protocol (ARP), Internet Control Message Protocol (ICMP), Internet Protocol version (IPv), and Logical Link Control (LLC).

Numeric features were scaled using MinMaxScaler to improve the model convergence and comparability across perturbation levels.

Machine Learning Models

Three machine learning models were employed for intrusion detection: XGBoost, random forest, and a MLP. The XGBoost classifier was configured with 100 estimators, a log-loss evaluation metric, and a fixed random seed for reproducibility, whereas the random forest model used 100 trees with the Gini impurity criterion. The MLP architecture consists of three hidden layers with 64 neurons each, utilizing Rectified Linear Unit (ReLU) activation and the Adam optimizer with a batch size of 128. All the models were trained using an 80/20 train–test split. To mitigate the effects of class imbalance, weighted loss functions were applied to the tree-based models. The model performance was evaluated using standard classification metrics, including precision, recall, F1-score, accuracy, and confusion matrices.

Table 1. Label distribution for binary classification.

Label	Count
Attack (1)	1,024,099
Benign (0)	24,476

Adversarial Perturbation Testing

To simulate adversarial attempts, the numerical features were perturbed using uniform noise. The perturbation levels were set at 5%, 10%, and 25%.

$$Xade[f] = X[f] \times (1 + \epsilon), \epsilon \sim U(-\epsilon, \epsilon)$$

Where, Xadv is the perturbed test set, and f represents numeric features. Recall was chosen as the primary metric to assess model sensitivity, as the detection of attacks is more critical than misclassifying benign flows.

RESULTS

Original ML Performance

On the original unperturbed test dataset, both XGBoost and random forest demonstrated strong detection capabilities, achieving consistently high recall for attack traffic, as summarized in Table 2. Their performance indicated the effective separation of malicious and benign flows despite the highly imbalanced class distribution. In contrast, the MLP model exhibited a degenerate prediction behavior, classifying nearly all samples as attacks and failing to identify benign traffic. This outcome highlights the sensitivity of neural-network-based models to severe class imbalance in intrusion detection tasks and underscores the practical advantage of ensemble tree-based methods for wireless traffic classification.

Feature Importance Analysis

Figures 1 and 2 illustrate the top 15 features contributing to the prediction performance of the XGBoost and random forest models. Both classifiers consistently rank TCP flag-based features such as ‘rst_count,’ ‘urg_count,’ and ‘syn_flag_number’ among the most influential. These features capture abnormal connection resets, urgency signaling, and connection initiation behavior, which are commonly observed in malicious traffic patterns such as scanning and denial-of-service attempts. The strong reliance on protocol-level indicators suggests that deviations in the standard TCP behavior remain a reliable signal for wireless intrusion detection across different learning paradigms.

In addition to protocol flags, both models emphasize statistical and timing-related flow characteristics, including variance, magnitude, IAT, and flow duration. These features reflect traffic burstiness and temporal irregularities that are difficult for adversaries to conceal without altering attack functionality. While XGBoost assigns a larger proportion of importance to a small subset of highly discriminative features, random forest distributes importance more evenly across multiple traffic statistics. This shared focus on aggregate behavioral features helps explain the observed robustness under adversarial perturbations because small feature-level modifications do not substantially change the underlying traffic dynamics used for detection.

Perturbation Testing

To examine the robustness of the model under realistic traffic variations, controlled perturbations were applied to the selected numerical features in the test set. These perturbations simulate benign fluctuations and adversarial manipulations commonly observed in wireless and IoT networks, such as timing instability, rate variation, and flow aggregation noise. Perturbation magnitudes of 5%, 10%, and 25% were evaluated to represent increasing levels of distortion while preserving the semantic validity of the network flows. The evaluation focused on attack-class recall because missed detections pose a higher operational risk than false alarms in IDS. The resulting performance values for each perturbation level are listed in Table 3.

As shown in Table 3, both the XGBoost and random forest models consistently maintained high attack recall across all perturbation levels. XGBoost exhibits negligible variation in recall, even at 25% perturbation, indicating that its decision boundaries are not overly sensitive to moderate feature fluctuations. A similar trend is observed for random forest, where recall remains stable and, in some cases, marginally improves under perturbation. This behavior suggests that ensemble-based tree models effectively rely on aggregated feature interactions rather than isolated values, rendering them resilient to noise. These results directly address the research question by demonstrating that commonly used machine learning models for wireless intrusion detection can remain robust when subjected to realistic adversarial perturbations in traffic features.

Table 2. Original model performance.

Model	Precision (0/1)	Recall (0/1)	F1-score (0/1)	Accuracy
XGBoost	0.79 / 1.00	0.998 / 0.994	0.88 / 0.997	0.9938
Random forest	0.84 / 1.00	0.993 / 0.996	0.91 / 0.997	0.9955
Multi-Layer Perceptron (MLP)	0.0 / 0.98	0.0 / 1.00	0.0 / 0.988	0.9767

Table 3. Recall under perturbations.

Model	ε = 0 %	ε = 5 %	ε = 10 %	ε = 25 %
XGBoost	0.9937	0.9948	0.9948	0.9948
Random forest	0.9956	0.9957	0.9957	0.9958

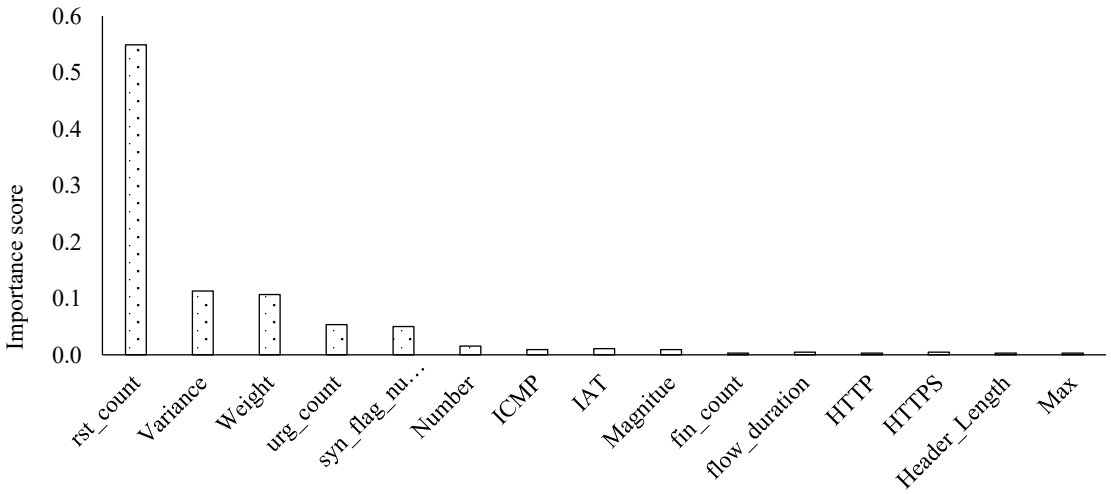

Figure 1. Top 15 feature importance (XGBoost).

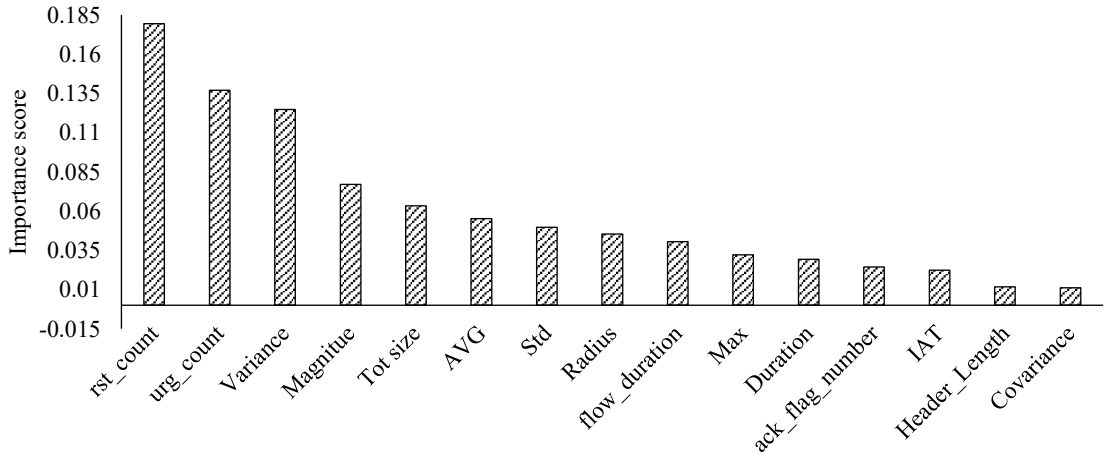

Figure 2. Top 15 feature importance (random forest).

DISCUSSION

The current results demonstrate that tree ensembles are resilient to adversarial perturbations in IoT traffic, whereas neural networks are vulnerable to imbalanced datasets. Feature importance indicates that timing and protocol flags are crucial for detecting attack flows, supporting findings from prior research [10–14].

The perturbation testing methodology provides a practical framework to evaluate ML robustness for deployment in real-world IDS, where attackers may subtly manipulate traffic. These limitations include single-dataset perturbations, binary classification, and synthetic perturbation models. Future work may extend to multiclass adversarial attacks, adaptive attackers, and online learning scenarios.

CONCLUSION

This study evaluated the robustness of commonly used machine learning models for wireless intrusion detection by examining their behavior under realistic adversarial perturbations applied to traffic features. Through controlled perturbation experiments, we observed that ensemble-based models, particularly XGBoost and random forest, consistently retained high attack recall even when feature values were altered by up to 25%. The minimal degradation in performance indicates that these models rely on stable and discriminative patterns within the network traffic rather than brittle feature thresholds. In contrast, the MLP demonstrated poor generalization in this setting, primarily due to severe class imbalance, underscoring the limitations of neural models when trained on skewed intrusion datasets without extensive balancing or regularization strategies.

Further analysis of the feature importance revealed that timing-related attributes and protocol-level indicators play a dominant role in robust intrusion detection. Features such as flow duration, IAT, and TCP flag counts were consistently ranked among the most influential across both tree-based models, suggesting that temporal dynamics and control-plane behavior are more resilient to perturbation than raw size-based statistics. These findings reinforce the need for IDS that emphasize behavior-driven features and robustness evaluation rather than relying solely on accuracy under clean conditions. Overall, this study highlights the importance of perturbation-aware testing when assessing intrusion detection models and provides practical insights for designing dependable IDS solutions in real-world IoT and wireless network environments.

Declaration of Interest

The authors declare no conflicts of interest regarding the publication of this manuscript.

Acknowledgement

We thank the Kaggle community for the IoT intrusion dataset used in this study.

REFERENCES

1. Al-Garadi MA, Mohamed A, Al-Ali AK, Du X, Ali I, Guizani M. A survey of machine and deep learning methods for internet of things (IoT) security. IEEE Commun Surv Tutor. 2020;22(3):1646–1685. doi:10.1109/COMST.2020.2988293.
2. Kumari P, Jain AK. A comprehensive study of DDoS attacks over IoT network and their countermeasures. Comput Secur. 2023;127:103096. doi:10.1016/j.cose.2023.103096.
3. Bankó MB, Dyszewski S, Králová M, Limpek MB, Papaioannou M, Choudhary G, et al. Advancements in machine learning-based intrusion detection in IoT: research trends and challenges. Algorithms. 2025;18(4):209. doi:10.3390/a18040209.
4. Liu N, Li C, Wang G, Wu Z, Li D. A dense mapping algorithm based on spatiotemporal consistency. Sensors (Basel). 2023;23(4):1876. doi:10.3390/s23041876.
5. Sharon Y, Berend D, Liu Y, Shabtai A, Elovici Y. TANTRA: timing-based adversarial network traffic reshaping attack. IEEE Trans Inf Forensics Secur. 2022;17:3225–3237. doi:10.1109/TIFS.2022.3201377.

6. Ibrahim Adamu A, Kumar Donta P, Mohd Ali D, Sarang S, Stojanović GM, Seroja Sarnin S. A systematic literature review of advanced machine learning techniques in wireless body area networks: application, challenges, and future directions. IEEE Access. 2025;13:194729–194778. doi:10.1109/ACCESS.2025.3631230.
7. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [preprint]. 2015. arXiv:1412.6572. doi:10.48550/arXiv.1412.6572.
8. Neto ECP, Dadkhah S, Ferreira R, Zohourian A, Lu R, Ghorbani AA. CICIoT2023: a real-time dataset and benchmark for large-scale attacks in IoT environment. Sensors (Basel). 2023;23(13):5941. doi:10.3390/s23135941.
9. Cyber Cop. (2023). IoT Intrusion Detection. [online] Kaggle.com. Available from: https://www.kaggle.com/datasets/subhajournal/iotintrusion
10. Alotaibi A, Rassam MA. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. Future Internet. 2023;15(2):62. doi:10.3390/fi15020062.
11. Harbi Y, Medani K, Gherbi C, Aliouat Z, Harous S. Roadmap of adversarial machine learning in Internet of Things-enabled security systems. Sensors. 2024;24(16):5150. doi:10.3390/s24165150.
12. Jamiri H, Zyane A. Adversarial attacks in IoT: A performance assessment of ML and DL models. Eng Proc. 2025;112(1):15. doi:10.3390/engproc2025112015.
13. Vitorino J, Praça I, Maia E. Towards adversarial realism and robust learning for IoT intrusion detection and classification. Ann Telecommun (Paris). 2023;78:401–412. doi:10.1007/s12243-023-00953-y.
14. Almousa O, Hamdallh B, Al-nu'man R. Enhancing IoT security: A comparative analysis of machine learning and deep learning techniques for botnet detection. Eng Technol Appl Sci Res. 2025;15(4):24498–24505. doi:10.48084/etasr.11092.

