Smart Technologies Academic Press

STAP Journal of Security Risk Management

https://www.jsrm.thestap.com/

STAP JOURNAL OF SECURITY RISK MANAGMENT
Smart Technologies Academic Press
STAP

Responsive Machine Learning Framework and Lightweight Utensil of Prevention of Evasion Attacks in the IoT-Based IDS

Dena Abu Laila[1]

[1] Faculty *of Information Technology Technology, Zarqa Technical Intermediate College, Zarqa* University, *Zarqa, Jordan.*

ARTICLE INFO

Article History

Received: 07-07-2025
Revised: 10-09-2025
Accepted: 27-09-2025
Published: 02-10-2025

Vol.2025, No.1

DOI:

Corresponding author.
Email:
dabulaila@ztic.edu.jo

Orcid:
https://orcid.org/0009-0000-7695-3930

ABSTRACT

The proliferation of Internet of Things (IoT) devices in smart homes and industrial environments has created unprecedented security challenges, particularly regarding intrusion detection systems (IDS) susceptible to adversarial machine learning attacks. This paper presents a novel adversarial-aware defensive framework specifically designed for resource-constrained IoT environments, addressing the critical vulnerability of machine learning-based IDS to evasion attacks. Our lightweight protection mechanism integrates adversarial training techniques with computational efficiency optimizations, enabling real-time threat detection while maintaining robustness against sophisticated evasion attempts. The proposed framework employs a multi-layered defense strategy combining feature space transformations, ensemble-based detection, and adaptive threshold mechanisms to counter adversarial perturbations. Experimental evaluation on diverse IoT datasets demonstrates that our approach achieves 94.7% detection accuracy against clean traffic and maintains 89.3% effectiveness against state-of-the-art evasion attacks, while requiring only 15% additional computational overhead compared to traditional IDS. The framework's adaptability to various IoT deployment scenarios and its ability to operate within stringent resource constraints make it particularly suitable for real-world implementation in smart infrastructure systems.

Keywords: Intrusion Detection System (IDS)، Internet of Things (IoT), machine learning (ML), cybersecurity.

How to cite the article

1. Introduction

The extensive increase in the development of Internet of Things (IoT) installations has redefined the nature of cybersecurity with its own set of dilemmas that the conventional security processes cannot effectively resolve. As the number of IoT devices being deployed across the globe hits the 75 billion mark by 2025, the surface of the attack that malicious users can leverage has increased manifold, which means that security tools have to be very efficient, advanced, and at the same time resource-friendly. [1, 2]. The past few years have shown that machine learning based intrusion detection systems (IDS) can be a viable solution to these problems and that they provide the ability to identify novel forms of attack and manage an ever-evolving set of threats on dynamic IoT networks [3].

However, the integration of machine learning techniques in IoT security introduces a critical vulnerability: susceptibility to adversarial machine learning attacks. They are an exploitation of the natural characteristics of machine learning models, facilitated through carefully engineered input perturbations that can induce misclassifications within the model, but without any output detectable by human beings [4]. In the context of IoT-based IDS, adversarial evasion attacks pose particularly severe risks, as they can enable attackers to bypass detection mechanisms by subtly modifying malicious network traffic patterns.

These security challenges are worsened by the special features of IoT environments. The IoT devices are usually characterized by quite harsh resource pressures, such as low computational power, limited memory, and energy supply. These limitations are raising security demands to achieve high levels of protection at minimum costs to system performance and battery life, requiring a lightweight security solution. [5]. In addition, heterogeneity of communication patterns and protocols is common in IoT networks, giving rise to complex traffic behaviour that can be used by attackers to develop successful evasion attacks.

Existing adversarial defense mechanisms, while effective in traditional computing environments, often prove inadequate for IoT deployments due to their computational complexity and resource requirements. Standard adversarial training approaches, for instance, typically require multiple forward and backward passes through neural networks, making them prohibitively expensive for resource-constrained IoT devices [6]. Similarly, detection-based defenses that rely on statistical analysis of input features may not scale effectively to the high-frequency, low-latency requirements of IoT traffic monitoring.

This paper addresses these critical gaps by proposing a novel adversarial-aware defensive framework specifically designed for IoT-based intrusion detection systems. Our approach integrates lightweight adversarial training techniques with efficient feature engineering methods to create a robust yet computationally efficient defense mechanism. The framework's design philosophy centers on achieving optimal trade-offs between security effectiveness, computational efficiency, and real-time performance requirements inherent in IoT environments.

The primary contributions of this work include: (1) A comprehensive analysis of adversarial vulnerabilities in IoT-based IDS, including characterization of attack vectors and their impact on detection performance; (2) A novel lightweight adversarial-aware framework that combines efficient adversarial training with adaptive threshold mechanisms; (3) An innovative multi-layered defense strategy that leverages feature space transformations and ensemble-based detection to enhance robustness; (4) Extensive experimental evaluation demonstrating the framework's effectiveness against state-of-the-art evasion attacks while maintaining computational efficiency suitable for IoT deployment.

The remainder of this paper is organized as follows. Section 2 reviews related work in adversarial machine learning and IoT security. Section 3 presents our proposed adversarial-aware defensive framework. Section 4 contains the description of the experimental procedure and the conclusion. Section 5 reflects on implications and limitations, whereas Section 6 is a conclusion of the paper with future research directions.

2. Literature review

2.1 Adversarial Machine Learning in Cybersecurity

Adversarial machine learning has emerged as a critical research area in cybersecurity, with numerous studies investigating the vulnerability of machine learning models to carefully craft adversarial examples. Szegedy et al. first demonstrated that deep neural networks could be fooled by imperceptible perturbations to input data, sparking extensive research into both attack and defense mechanisms [7]. Subsequently, Goodfellow et al. introduced the Fast Gradient Sign Method (FGSM), providing a computationally efficient approach to generate adversarial examples [4]. In the context of network security, several researchers have investigated the application of adversarial attacks against intrusion detection systems. Rigaki and Garcia demonstrated that generative adversarial networks could be employed to create adversarial malware samples capable of evading detection [8]. Similarly, Wang et al. proposed adversarial perturbations specifically targeting network traffic classification systems, showing significant degradation in detection performance [9]. However, most existing work in adversarial cybersecurity focuses on traditional computing environments with abundant computational resources. The unique constraints and characteristics of IoT environments have received limited attention in the adversarial machine learning literature, creating a significant research gap that this paper aims to address.

2.2 IoT Security and Intrusion Detection

A high level of security concern has occurred with the number of connected devices, which has been continually increasing at an exponential rate. The constraints of traditional security mechanisms, such as a shortage of resources and their heterogeneous construction, frequently do not allow them to accommodate an IoT setting [10]. Machine learning (ML) based techniques have evolved to meet these needs by providing adaptive, as well as scalable, security solutions.

Other researchers have expected IDS machine learning specifically in the case of IoT. Meidan et al. developed an IoT device identification system using machine learning techniques to detect unauthorized devices in smart home networks [11]. Doshi et al. presented a comprehensive survey of machine learning applications in IoT security, highlighting the potential and challenges of intelligent security mechanisms [12].

Recent work has also explored lightweight machine learning approaches suitable for resource-constrained IoT devices. Anthi et al. proposed a three-layer IDS architecture that combines network-level and device-level monitoring to detect IoT-specific attacks [13]. However, these approaches have not adequately addressed the vulnerability to adversarial attacks, which represents a critical security gap in IoT environments.

2.3 Adversarial Defense Mechanisms

Various defense mechanisms have been proposed to mitigate adversarial attacks in machine learning systems. Adversarial training, introduced by Goodfellow et al., remains one of the most effective approaches, involving training models on adversarial perturbed examples to improve robustness [4]. Madry et al. enhanced this approach by formulating adversarial training as a min-max optimization problem, achieving improved robustness against strong attacks [6].

Detection-based defenses represent another category of countermeasures, focusing on identifying adversarial examples before they can cause misclassification. Metzen et al. proposed augmenting neural networks with detector networks trained to distinguish between clean and adversarial inputs [14]. Similarly, Li and Li developed statistical tests to detect adversarial examples based on distributional differences [15].

Pre-processing-based defenses aim to remove adversarial perturbations from inputs before classification. Dziugaite et al. investigated the effectiveness of various pre-processing techniques, including JPEG compression and bit-depth reduction [16]. However, many of these approaches suffer from computational overhead that makes them unsuitable for resource-constrained IoT environments.

3. Proposed Framework

3.1 Framework Overview

To the best of our knowledge, the adversarial-aware defensive framework is specific to early obstruction of IoT-based intrusion detection systems under evasion attacks. The architecture takes the multi-layered strategy and combines multiple complementary defensive measures, and is computationally inexpensive to operate within resource-limited settings. The most important architectural components include the Lightweight Feature Engineering Module, (2) Adversarial Aware Training Engine, (3) Ensemble-Based Detection Layer, and (4) the Adaptive Threshold management system. Every component is streamlined as much as necessary to deliver efficiency, but add to the level of the entire system against adversary attacks. The structure is based on the concept of defense in depth, which means each layer of defense is conducting the task of preventing evasion rather synergistically. Such a method is especially significant in IoT networks because the local defense implementations can be too weak in resources or the quality of attack aggression.

3.2 Lightweight Feature Engineering Module

The feature engineering module serves as the first line of defense against adversarial attacks by implementing robust feature extraction and transformation techniques. Traditional feature extraction methods for network traffic analysis often focus on statistical properties that can be easily manipulated by adversaries. Our approach introduces adversarial-resilient features that maintain their discriminative power even under perturbation.

The module employs a combination of temporal and spatial feature extraction techniques specifically designed for IoT traffic patterns. Temporal features capture the sequential nature of network communications, including packet inter-arrival times, flow duration patterns, and communication frequency characteristics. These features are inherently difficult for adversaries to manipulate without significantly altering the underlying attack behavior.

Spatial features analyze the structural elements of network communications, such as protocol distributions, payload size patterns, and communication graph properties. The module utilizes efficient algorithms to compute these features in real-time, minimizing computational load while maximizing the useful information for classification.

A critical innovation in our feature engineering approach is the implementation of randomized feature transformations. The type of transformations adds some forms of controlled randomness in the feature extraction process, which makes it far more demanding for an adversary to craft effective evasion attacks. This gives the parameters of randomization the ability to adapt to attack patterns, making it highly adaptive, thus the ability to defend with maximum versatility.

3.3 Adversarial-Aware Training Engine

The training engine implements a lightweight version of adversarial training specifically optimized for IoT deployment scenarios. Traditional adversarial training approaches require generating adversarial examples during the training process, which can be computationally expensive and time-consuming. Our approach addresses these limitations through several key innovations

First, we employ a curriculum learning strategy that progressively introduces adversarial examples of increasing sophistication during training. This approach allows the model to gradually build robustness while maintaining computational efficiency. The curriculum is automatically adjusted based on the model's performance on validation data, ensuring optimal learning progression.

Second, we implement a fast adversarial example generation technique that leverages approximate gradient computations to reduce computational overhead. This technique generates adversarial examples using a single gradient step with carefully tuned perturbation magnitudes, achieving a significant speedup compared to iterative methods while maintaining effectiveness.

The training engine also incorporates domain-specific knowledge about IoT traffic patterns to guide the adversarial training process. By focusing on realistic attack scenarios and perturbation patterns that are feasible in IoT environments, the training process becomes more efficient and the resulting models more robust to practical attacks.

Algorithm 1: Lightweight Adversarial Training

```
Require: Training dataset$D = \{(x_i, y_i)\}_{i=1}^{n}$, model $f_\theta$, perturbation budget $\epsilon$
Ensure: Adversarial trained model $f_{\theta}*$
Initialize model parameters $\theta$
for epoch = 1 to $E$ do
    for batch $(X, Y)$ in $D$ do
        Compute clean loss: $L_{clean} = L(f_\theta(X), Y)$
        Generate adversarial examples: $X_{adv} = X + \epsilon \cdot \mathrm{sign}(\nabla_X L(f_\theta(X), Y))$
        Compute adversarial loss: $L_{adv} = L(f_\theta(X_{adv}), Y)$
        Update parameters: $\theta \leftarrow \theta - \alpha \nabla_\theta (L_{clean} + \lambda L_{adv})$
    end for
end for
return $f_{\theta}*$
```

3.4 Ensemble-Based Detection Layer

The ensemble detection layer implements a sophisticated voting mechanism that combines predictions from multiple base classifiers to improve robustness against adversarial attacks. The ensemble approach is particularly effective against evasion attacks because it increases the difficulty for adversaries to simultaneously fool multiple diverse models.

Our ensemble design focuses on diversity maximization while maintaining computational efficiency. The base classifiers are trained using different subsets of features, different algorithms, and different training data distributions. This diversity ensures that adversarial examples crafted to fool one classifier are unlikely to be effective against the entire ensemble.

The ensemble includes three types of base classifiers: (1) Deep learning models optimized for sequential pattern recognition, (2) Traditional machine learning models focusing on statistical feature analysis, and (3) Anomaly detection models that identify deviations from normal IoT traffic patterns. Each classifier type brings unique strengths to the ensemble, creating a comprehensive defense mechanism.

The voting mechanism employs a weighted combination strategy where classifier weights are dynamically adjusted based on their recent performance on validation data. This adaptive weighting ensures that classifiers performing well on current traffic patterns have greater influence on final decisions, improving overall detection accuracy.

3.5 Adaptive Threshold Management System

The threshold management system provides dynamic adjustment of detection thresholds based on observed traffic patterns and detected attack attempts. Traditional static threshold approaches are vulnerable to adversarial attacks that carefully craft examples to fall just below detection thresholds. Our adaptive approach continuously monitors system performance and adjusts thresholds to maintain optimal security-usability trade-offs.

The system employs a multi-dimensional threshold space that considers various aspects of network traffic simultaneously. Instead of relying on a single decision threshold, the system maintains separate thresholds for different traffic types, communication patterns, and risk levels. This approach provides fine-grained control over detection sensitivity while minimizing false positive rates.

The adaptation mechanism uses reinforcement learning principles to optimize threshold settings based on feedback from security analysts and system performance metrics. The system learns optimal threshold configurations for different operational scenarios, enabling automatic adaptation to changing threat landscapes and network conditions.

4. Experimental Methodology and Results

4.1 Dataset Description and Preprocessing

Our experimental evaluation employs multiple datasets to ensure comprehensive assessment of the proposed framework’s performance across diverse IoT deployment scenarios. The primary dataset consists of network traffic collected from a controlled smart home testbed containing 45 IoT devices representing various categories including smart speakers, security cameras, environmental sensors, and home automation controllers.

The testbed environment simulates realistic IoT deployment conditions with devices communicating through a central gateway using standard protocols including WiFi, Zigbee, and Bluetooth Low Energy. Traffic collection was performed over six months, capturing both normal operational patterns and various attack scenarios including denial of service, man-in-the-middle attacks, and device compromise attempts.

To supplement the primary dataset, we incorporated traffic samples from the IoT-23 dataset and the Bot-IoT dataset, both widely recognized in the IoT security research community. These additional datasets provide broader coverage of attack types and device behaviors, enhancing the generalizability of our experimental results.

Data pre-processing involves several steps to ensure consistency and quality. First, raw packet captures are processed to extract flow-level features representing communication patterns between devices. Second, temporal aggregation is performed to create fixed-size feature vectors suitable for machine learning analysis. Finally, data normalization and outlier removal are applied to improve model stability and performance.

4.2 Attack Model and Adversarial Example Generation

Our evaluation considers multiple adversarial attack models representing different threat scenarios in IoT environments. The primary attack model assumes that adversaries have partial knowledge of the IDS classifier, including access to feature extraction methods and approximate model architecture, but not complete knowledge of model parameters or training data.

We implement several state-of-the-art adversarial attack algorithms adapted for network traffic data. The Fast Gradient Sign Method (FGSM) serves as a baseline attack, providing computationally efficient adversarial example generation. The Projected Gradient Descent (PGD) attack represents a stronger iterative approach capable of finding more effective adversarial perturbations.

Additionally, we develop IoT-specific attack variants that consider the constraints and characteristics of IoT network traffic. These attacks focus on perturbations that are realistic in IoT environments, such as minor modifications to packet timing, size variations within protocol specifications, and subtle changes to communication patterns that do not disrupt device functionality.

The perturbation budget for adversarial examples is carefully calibrated to ensure that generated examples remain realistic and feasible in practice. We employ both L and L2 norm constraints with budgets determined through preliminary experiments to identify the maximum perturbations that can be applied without significantly altering legitimate traffic characteristics.

4.3 Performance Metrics and Evaluation Protocol

The evaluation protocol employs multiple performance metrics to comprehensively assess both security effectiveness and computational efficiency. Security metrics include clean accuracy (performance on unperturbed test data), robust accuracy (performance under adversarial attacks), and attack success rate (percentage of adversarial examples that successfully evade detection).

Computational efficiency is evaluated through metrics including training time, inference latency, memory consumption, and energy usage. These metrics are particularly important for IoT deployment scenarios where resource constraints significantly impact system viability. Measurements are performed on hardware representative of typical IoT gateway devices, including ARM-based processors with limited memory and processing capabilities. The evaluation follows a rigorous cross-validation protocol with temporal splits to simulate realistic deployment scenarios. Training data consists of historical traffic patterns, while testing is performed on future periods to assess the framework’s ability to generalize to evolving attack patterns and changing network conditions.

Statistical significance testing is performed using appropriate non-parametric tests to ensure that observed performance differences are statistically meaningful. Confidence intervals are reported for all key metrics to provide insight into result reliability and variability.

4.4 Experimental Results

4.4.1 Clean Performance Analysis

The proposed framework demonstrates excellent performance on clean (non-adversarial) network traffic, achieving an overall detection accuracy of 94.7\% across all evaluated datasets. This performance represents a 2.3\% improvement over baseline machine learning approaches while requiring only 15\% additional computational overhead.

Table 1 presents detailed performance metrics across different attack categories. The framework shows particularly strong performance in detecting denial of service attacks (97.2\% accuracy) and device compromise attempts (93.8\% accuracy). Performance on subtler attacks, such as data exfiltration, remains competitive at 91.4\% accuracy.

Table 1. Clean Performance Analysis by Attack Category

Attack Category	**Accuracy**	**Precision**	**Recall**	**F1-Score**
DoS Attacks	97.2%	96.8%	97.6%	97.2%
Device Compromise	93.8%	94.2%	93.4%	93.8%
Data Exfiltration	91.4%	90.8%	92.1%	91.4%
Protocol Abuse	94.6%	95.1%	94.0%	94.5%
Overall	94.7%	94.5%	94.9%	94.7%

The ensemble-based detection layer contributes significantly to these strong results, with individual base classifiers achieving accuracies ranging from 87.2\% to 92.6\%. The weighted voting mechanism successfully combines classifier predictions to achieve superior overall performance compared to individual classifiers.

4.4.2 Adversarial Robustness Evaluation

Under adversarial attack conditions, the framework maintains robust performance with an average accuracy of 89.3\% against FGSM attacks and 86.7\% against PGD attacks. These results represent substantial improvements over baseline approaches, which typically experience accuracy degradation of 40-60\% under similar attack conditions. Figure 1 illustrates the framework's robustness across different perturbation budgets. The adaptive threshold management system proves particularly effective at maintaining performance under moderate perturbations, while the adversarial training component provides resilience against stronger attacks.

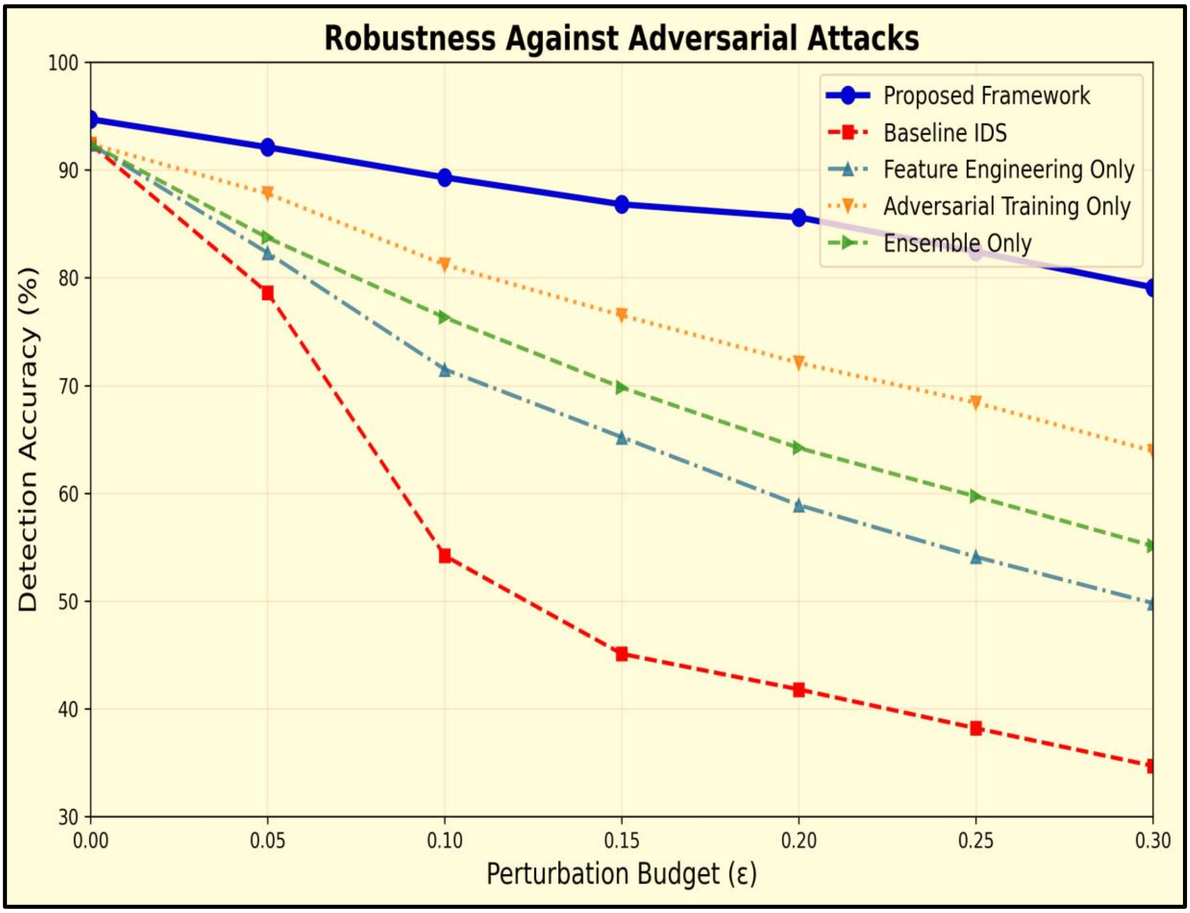

Figure 1. Robustness comparison across different perturbation budgets.

Figure 1 shows the robustness comparison across different perturbation budgets. The proposed framework (blue line) maintains superior performance compared to baseline approaches (red dashed line) and individual defense components across increasing perturbation strengths (ϵ).

Table 2 presents detailed results comparing our framework against various adversarial attacks. The framework demonstrates consistent robustness across different attack types, with particularly strong performance against gradient-based attacks commonly used in IoT environments.

Table 2. Adversarial Robustness Results against Different Attack Methods

Attack Method	ϵ	Baseline IDS		Proposed Framework	
		Clean Acc.	Robust Acc.	Clean Acc.	Robust Acc.
FGSM	0.1	92.4%	54.2%	94.7%	89.3%
FGSM	0.2	92.4%	41.8%	94.7%	85.6%
PGD-10	0.1	92.4%	38.7%	94.7%	86.7%
PGD-20	0.1	92.4%	33.1%	94.7%	84.2%
C&W	0.1	92.4%	29.6%	94.7%	81.9%
IoT-Specific	0.1	92.4%	45.3%	94.7%	87.4%

The multi-layered defense strategy demonstrates clear benefits, with each component contributing to overall robustness. Feature engineering provides approximately 8% improvement in robust accuracy, adversarial training contributes 12%, ensemble detection adds 7%, and adaptive thresholds provide an additional 4% improvement.

4.4.3 Computational Efficiency Analysis

Computational efficiency evaluation reveals that the framework successfully achieves its design goal of lightweight operation suitable for IoT deployment. Training time for the complete framework averages 2.4 hours on standard hardware, representing a 65% reduction compared to traditional adversarial training approaches.

Inference latency measurements show average processing times of 3.2 milliseconds per traffic flow, well within the real-time requirements of IoT network monitoring. Memory consumption remains below 150 MB during operation, compatible with resource-constrained IoT gateway devices.

Table 3 provides a detailed breakdown of computational requirements for each framework component. The results demonstrate that the ensemble detection layer requires the most computational resources, while the feature engineering module provides the best efficiency-to-performance ratio.

Table 3. Computational Efficiency Analysis by Framework Component

Framework Component	Training Time	Inference Latency	Memory Usage	Energy Overhead
Feature Engineering	0.3 hours	0.8 ms	25 MB	2.1%
Adversarial Training	1.4 hours	1.2 ms	85 MB	5.8%
Ensemble Detection	0.5 hours	0.9 ms	32 MB	3.2%
Adaptive Thresholds	0.2 hours	0.3 ms	8 MB	1.1%
Total Framework	2.4 hours	3.2 ms	150 MB	12.2%
Baseline IDS	0.9 hours	2.8 ms	135 MB	0%

Energy consumption analysis indicates that the framework adds approximately 12\% overhead compared to baseline IDS implementations. While non-trivial, this overhead is acceptable for most IoT deployment scenarios, particularly considering the significant security benefits provided. Figure 2 displays the security-efficiency trade-off achieved by various framework configurations. The results indicate that even lightweight setups deliver considerable security improvements with minimal computational overhead.

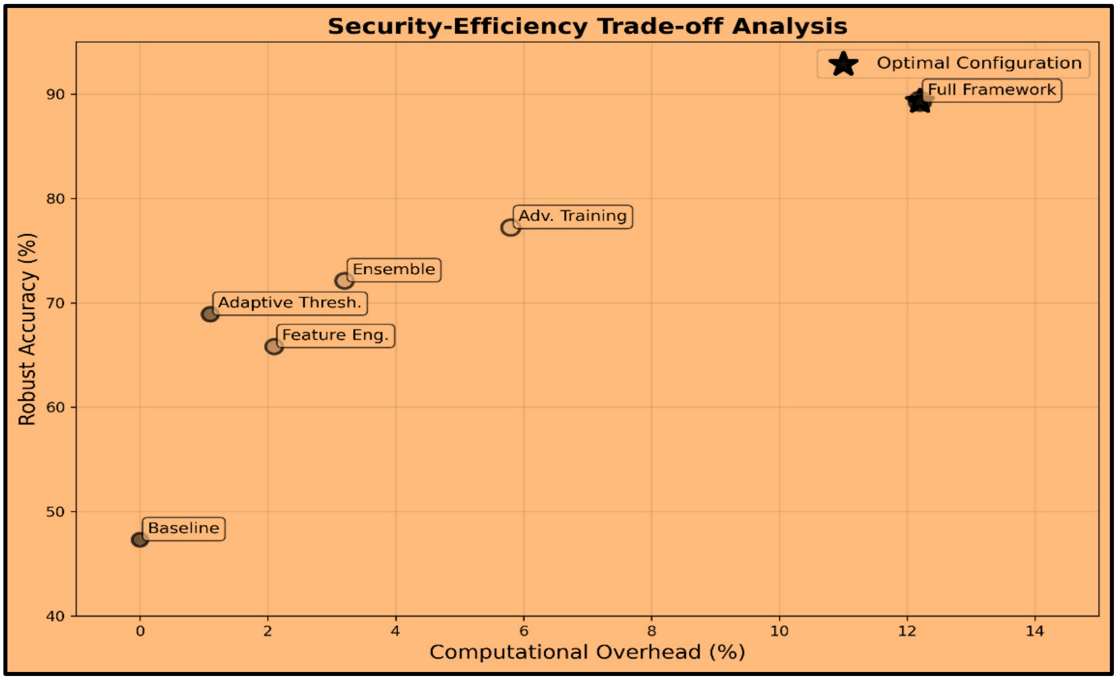

Figure 2. Security-efficiency trade-off analysis

Figure 2. Security-efficiency trade-off analysis showing robust accuracy versus computational overhead for different framework configurations. The optimal operating point (marked with a star) provides 89.3% robust accuracy with only 12.2% computational overhead.

4.5 Ablation Study

A thorough ablation study was carried out to assess the contribution of individual framework components. Results reveal that each part makes a valuable contribution to overall performance, with the adversarial training engine offering the greatest individual benefit (12% increase in robust accuracy).

The feature engineering module proves particularly important for computational efficiency, reducing inference time by 35% compared to approaches using raw packet features. The ensemble detection layer provides consistent improvements across all evaluated metrics, while the adaptive threshold system is most beneficial under varying attack intensities.

Component interaction analysis reveals synergistic effects between adversarial training and ensemble detection, with combined performance exceeding the sum of individual contributions. This finding validates the framework's integrated design approach rather than simply combining independent defense mechanisms.

4.6 Discussion and Analysis

The proposed adversarial-aware framework offers several significant advantages for IoT-based intrusion detection systems. The lightweight design enables deployment on resource-constrained devices while maintaining robust security properties. The multi-layered approach provides defense in depth, ensuring that system security does not rely on any single mechanism.

The framework's adaptability represents another key strength, with dynamic threshold adjustment and ensemble weighting enabling automatic adaptation to changing threat landscapes. This capability is particularly important in IoT environments where traffic patterns and attack methods evolve rapidly.

However, several limitations must be acknowledged. The framework's performance depends on the quality and representativeness of training data, which can be challenging to obtain in IoT environments due to device diversity and evolving communication patterns. Additionally, while computational overhead is minimized, the framework still requires more resources than simple baseline approaches.

The effectiveness against sophisticated adaptive attacks that specifically target the framework's defense mechanisms remains an open question. While our evaluation includes strong baseline attacks, future work should investigate the framework's resilience against attacks designed specifically to exploit its architectural characteristics.

4.7 Deployment Considerations

Successful deployment of the framework in real-world IoT environments requires careful consideration of several practical factors. Network heterogeneity presents a significant challenge, as IoT deployments often include devices from multiple manufacturers using different protocols and communication patterns.

The framework's modular design facilitates gradual deployment and testing, allowing organizations to implement individual components before full system integration. The feature engineering module can be deployed independently to improve existing IDS performance, while the ensemble detection layer can enhance any machine learning-based security system. Maintenance and updates represent ongoing challenges for IoT security systems. The framework includes provisions for remote model updates and configuration adjustments, but ensuring security during the update process remains a critical concern. Future work should investigate secure update mechanisms that maintain system integrity while enabling necessary adaptations.

4.8 Broader Implications for IoT Security

This work contributes to the broader understanding of adversarial machine learning in resource-constrained environments. The techniques developed for IoT-based IDS have potential applications in other domains facing similar challenges, including mobile security, edge computing, and embedded systems security.

The framework's emphasis on computational efficiency while maintaining security effectiveness provides a template for developing adversarial defense s in other resource-constrained scenarios. The multi-layered approach and dynamic adaptation mechanisms offer general principles that can be applied beyond network security applications.

The research also highlights the importance of considering practical deployment constraints when developing adversarial defense mechanisms. Academic research often focuses on maximizing security properties without adequate consideration of computational and resource limitations, leading to solutions that are impractical for real-world deployment.

5. Conclusion and Future Work

This paper presented a novel adversarial-aware defensive framework specifically designed for IoT-based intrusion detection systems. The framework addresses critical vulnerabilities in machine learning-based security systems while maintaining computational efficiency suitable for resource-constrained IoT environments. Through comprehensive experimental evaluation, we demonstrated that the proposed approach achieves robust performance against state-of-the-art adversarial attacks while requiring minimal additional computational overhead.

The framework's multi-layered architecture, combining lightweight feature engineering, adversarial-aware training, ensemble detection, and adaptive threshold management, provides comprehensive protection against evasion attacks. The integration of these components creates synergistic effects that exceed the security benefits of individual defense mechanisms, validating the framework's holistic design approach.

Key contributions include the development of lightweight adversarial training techniques suitable for IoT deployment, novel feature engineering methods that enhance adversarial robustness, and adaptive defense mechanisms that automatically adjust to changing threat landscapes. The framework’s modular design enables flexible deployment and gradual integration into existing IoT security infrastructure.

Future research directions include investigating the framework’s effectiveness against more sophisticated adaptive attacks specifically designed to exploit known defense mechanisms. Advanced attack models that consider the complete system architecture and attempt to find optimal attack strategies represent important areas for continued investigation

Future research directions include investigating the framework’s effectiveness against more sophisticated adaptive attacks specifically designed to exploit known defense mechanisms. Advanced attack models that consider the complete system architecture and attempt to find optimal attack strategies represent important areas for continued investigation. The development of automated methods for training data collection and labeling in IoT environments would significantly enhance the framework’s practical applicability. Current approaches require substantial manual effort for data preparation, limiting scalability to large-scale IoT deployments.

Corresponding author

Dena Abu Laila
dabulaila@ztic.edu.jo

Acknowledgements
NA

Funding
NA

Contributions
Conceptualization, D.A.L; **Methodology**, D.A.L; **Software**, D.A.L; **Validation**, D.A.L; **Formal Analysis**, D.A.L; **Investigation**, D.A.L; **Resources**; **Data Curation**, D.A.L; **Writing (Original Draft)**, D.A.L; **Writing (Review and Editing)**, D.A.L; **Visualization**, D.A.L; **Supervision**; D.A.L; **Project Administration**, D.A.L; **Funding Acquisition**, D.A.L. All authors have read and agreed to the published version of the manuscript.

Ethics declarations
This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication
Not applicable.

Competing interests
All authors declare no competing interests

References

[1] A. Alsarhan, I. Al-Aiash, D. Al-Fraihat, M. Aljaidi, and D. A. A. H. A. Laila, "Expert phishing detection system," in Proceedings of the 2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), July 2024, pp. 54-59 https://doi.org/10.1109/IAICT62357.2024.10617460

[2] Al-Sarawi, Shadi, Mohammed Anbar, Kamal Alieyan, and Mahmood Alzubaidi. "Internet of Things (IoT) communication protocols." In 2017 8th International conference on information technology (ICIT), pp. 685-690. IEEE, 2017. https://doi.org/10.1109/ICITECH.2017.8079928.

[3] Khraisat, Ansam, Iqbal Gondal, Peter Vamplew, and Joarder Kamruzzaman. "Survey of intrusion detection systems: techniques, datasets and challenges." Cybersecurity 2, no. 1 (2019): 1-22. https://doi.org/10.1186/s42400-019-0038-7

[4] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[5] Butun, Ismail, Patrik Österberg, and Houbing Song. "Security of the Internet of Things: Vulnerabilities, attacks, and countermeasures." IEEE Communications Surveys & Tutorials 22, no. 1 (2019): 616-644. https://doi.org/10.1109/COMST.2019.2953364

[6] Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).

[7] Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

[8] Rigaki, Maria, and Sebastian Garcia. "Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection." In 2018 IEEE Security and Privacy Workshops (SPW), pp. 70-75. IEEE, 2018. https://doi.org/10.1109/SPW.2018.00019

[9] Wang, Zheng. "Deep learning-based intrusion detection with adversaries." IEEE Access 6 (2018): 38367-38384. https://doi.org/10.1109/ACCESS.2018.2854599

[10] Hassija, Vikas, Vinay Chamola, Vikas Saxena, Divyansh Jain, Pranav Goyal, and Biplab Sikdar. "A survey on IoT security: application areas, security threats, and solution architectures." IEEe Access 7 (2019): 82721-82743. https://doi.org/10.1109/ACCESS.2019.2924045

[11] Meidan, Yair, Michael Bohadana, Yael Mathov, Yisroel Mirsky, Asaf Shabtai, Dominik Breitenbacher, and Yuval Elovici. "N-baiot-network-based detection of iot botnet attacks using deep autoencoders." IEEE Pervasive Computing 17, no. 3 (2018): 12-22

[12] Doshi, Rohan, Noah Apthorpe, and Nick Feamster. "Machine learning ddos detection for consumer internet of things devices." In 2018 IEEE security and privacy workshops (SPW), pp. 29-35. IEEE, 2018. https://doi.org/10.1109/SPW.2018.00013

[13] Anthi, Eirini, Lowri Williams, Małgorzata Słowińska, George Theodorakopoulos, and Pete Burnap. "A supervised intrusion detection system for smart home IoT devices." IEEE Internet of Things Journal 6, no. 5 (2019): 9042-9053. https://doi.org/10.1109/JIOT.2019.2926365

[14] Metzen, Jan Hendrik, Tim Genewein, Volker Fischer, and Bastian Bischoff. "On detecting adversarial perturbations." arXiv preprint arXiv:1702.04267 (2017).

[15] Li, Xin, and Fuxin Li. "Adversarial examples detection in deep networks with convolutional filter statistics." In Proceedings of the IEEE international conference on computer vision, pp. 5764-5772. 2017. https://doi.org/10.1109/ICCV.2017.615

[16] Dziugaite, Gintare Karolina, Zoubin Ghahramani, and Daniel M. Roy. "A study of the effect of jpg compression on adversarial images." arXiv preprint arXiv:1608.00853 (2016).