DOI:doi.org/10.65766/alyj.2026.24.01.16

# Adversarial Machine Learning for Detection of Zero-Day Attacks on IoT Networks

Ahmed Gheni Dawood
College of Education for Humanities - University of Diyala – Iraq
Corresponding Author Email: Ahmed.hum@uodiyala.edu.iq

**Important Dates**
Received:2026/4/8, Accepted:2026/6/24 , Published: 30/6/2026



**Abstract**

The Internet of Things (IoT) has dramatically expanded global connectivity, with billions of heterogeneous devices generating massive volumes of network traffic. This growth has significantly enlarged the attack surface for cyber threats, particularly synthetic unknown attacks that exploit vulnerabilities not yet captured in existing signature databases. Traditional intrusion detection systems (IDS), which rely on known attack signatures, fail to detect such novel threats. This study proposes a comprehensive Adversarial Machine Learning (AML) framework designed to detect synthetic unknown attacks in IoT networks through the integration of three complementary techniques: Generative Adversarial Networks (GANs) for synthesizing realistic attack samples, Adversarial Autoencoders (AAEs) for robust low-dimensional feature extraction, and Deep Neural Networks (DNNs) for real-time anomaly classification. The framework was evaluated on a controlled synthetic IoT dataset comprising 250,000 traffic samples. Results demonstrate detection accuracy of 94.8%, precision of 93.1%, recall of 92.7%, and F1-score of 92.9%, substantially outperforming signature-based IDS (63.5%), SVM (87.2%), and Random Forest (90.1%). The framework maintained detection rates of 90.2%, 87.5%, and 85.3% under FGSM, PGD, and Carlini-Wagner adversarial perturbations, respectively. Scalability testing across datasets up to 4 million samples confirmed sustained performance with inference latency below 0.021 seconds per sample. The proposed framework represents a scalable, adversarially resilient approach to IoT intrusion detection, with limitations regarding real-world generalization openly acknowledged.

**Keywords:** Adversarial Machine Learning, IoT Security, Generative Adversarial Networks, Adversarial Autoencoders, Anomaly Detection, Real-Time Detection.

# استخدام تعلم الآلة الخصمي (أو التنافسي) للكشف عن هجمات الثغرات الصفرية في شبكات إنترنت الأشياء

**احمد غني داود**

كلية التربية للعلوم الإنسانية – جامعة ديالى – العراق

**ايميل الباحث المراسل:** Ahmed.hum@uodiyala.edu.iq

DOI:doi.org/10.65766/alyj.2026.24.01.16

**الملخص**

أدى إنترنت الأشياء (IoT) إلى توسع هائل في الاتصال العالمي، حيث تولد مليارات الأجهزة المتنوعة كميات هائلة من حركة مرور الشبكة. وقد أدى هذا النمو إلى زيادة كبيرة في مساحة الهجمات الإلكترونية، لا سيما الهجمات المصطنعة غير المعروفة التي تستغل الثغرات الأمنية غير المسجلة في قواعد بيانات التوقيعات الحالية. وتفشل أنظمة كشف التسلل التقليدية(IDS) ، التي تعتمد على توقيعات الهجمات المعروفة، في اكتشاف هذه التهديدات الجديدة. تقترح هذه الدراسة إطار عمل شامل للتعلم الآلي التنافسي (AML) مصمما لأكتشاف الهجمات المصطنعة غير المعروفة في شبكات إنترنت الأشياء من خلال دمج ثلاث تقنيات متكاملة: الشبكات التوليدية التنافسية (GANs) لتوليد عينات هجوم واقعية، والمشفرات التلقائية التنافسية (AAEs) لاستخراج ميزات قوية منخفضة الأبعاد، والشبكات العصبية العميقة (DNNs) لتصنيف الحالات الشاذة في الوقت الفعلي. تم تقييم إطار العمل على مجموعة بيانات اصطناعية مضبوطة لإنترنت الأشياء تضم 250,000 عينة من حركة المرور. أظهرت النتائج دقة كشف بلغت 94.8%، ودقة 93.1%، واستدعاء 92.7%، ودرجة F1 بلغت 92.9%، متفوقة بذلك بشكل ملحوظ على أنظمة كشف التسلل القائمة على التوقيعات (63.5%)، وخوارزمية(87.2%) SVM ، وخوارزمية الغابة العشوائية (90.1%). وحافظ الإطار على معدلات كشف بلغت 90.2%، و87.5%، و85.3% في ظل تشويشات FGSM، وPGD، و Carlini-Wagnerالمعادية، على التوالي. وأكد اختبار قابلية التوسع عبر مجموعات بيانات تصل إلى 4 ملايين عينة استمرار الأداء مع زمن استجابة للاستدلال أقل من 0.021 ثانية لكل عينة. يمثل الإطار المقترح نهجاً قابلاً للتوسع ومقاوماً للهجمات المعادية لكشف التسلل في إنترنت الأشياء، مع الإقرار بوجود قيود تتعلق بالتعميم في العالم الحقيقي.

**الكلمات المفتاحية:** التعلم الآلي التنافسي، أمن إنترنت الأشياء، الشبكات التنافسية التوليدية، المشفرات التلقائية التنافسية، اكتشاف الشذوذ، الاكتشاف في الوقت الحقيقي.

## 1. Introduction

The Internet of Things (IoT) has transformed modern infrastructure, with connected devices projected to exceed 40 billion by 2025, generating over 80 zettabytes of data annually [Al-Garadi, Guizani (2020)]. IoT ecosystems span smart home appliances, health wearables, industrial sensors, autonomous vehicles, and smart city infrastructure. This unprecedented expansion has simultaneously enlarged the attack surface for sophisticated cyber threats. Of particular concern are synthetic unknown attacks — threats exploiting vulnerabilities not yet documented in any signature database — against which conventional signature-based IDS are fundamentally ineffective, since they can only recognize threats after prior exposure [Alkadi, S., Al-Ahmadi, S., & Ismail, M. M. B. (2023)].

IoT devices are uniquely vulnerable due to hardware constraints (16-64 MB RAM), architectural heterogeneity (ARM, MIPS, RISC-V), and infrequent firmware updates, factors that compound the difficulty of deploying effective security mechanisms [Bhadauria & Sanyal(2021)]. Attack vectors include Mirai-style botnets exploiting default credentials, protocol-level attacks on lightweight standards such as MQTT and CoAP, firmware exploitation, and data poisoning against machine learning pipelines. Approximately 65% of IoT devices carry unpatched firmware vulnerabilities, with 25% of known attacks targeting firmware and 30% exploiting protocol weaknesses [Carlini & Wagner (2017)].

**Research Problem and Objectives:** The central challenge addressed in this work is the detection of synthetic unknown attacks in resource-constrained IoT network environments, while maintaining resilience against adversarial manipulation. The specific objectives are: (1) to design an integrated AML framework combining GANs, AAEs, and DNNs for IoT anomaly detection; (2) to evaluate the

DOI:doi.org/10.65766/alyj.2026.24.01.16

framework's detection accuracy, adversarial robustness, scalability, and real-time performance; and (3) to identify limitations and deployment considerations relevant to practical IoT security contexts.

Adversarial Machine Learning (AML) offers a principled approach by training models to remain robust against adversarial perturbations — subtle input modifications designed to evade detection [Kurakin, Goodfellow & Bengio (2017)]. This paper presents an AML framework evaluated on a controlled synthetic IoT dataset, with explicit acknowledgment of the distinction between simulated unknown attacks and verified real-world zero-day exploits. The paper is organized as follows: Section 2 reviews related work; Section 3 describes the dataset; Section 4 presents the models and tools; Section 5 details the proposed methodology; Section 6 reports experimental results; Section 7 provides discussion; Section 8 concludes the paper.

As shown in Figure 1, the parallel growth of IoT device deployment and cyber threat incidents between 2015 and 2025 underscores the urgency of developing scalable security solutions.

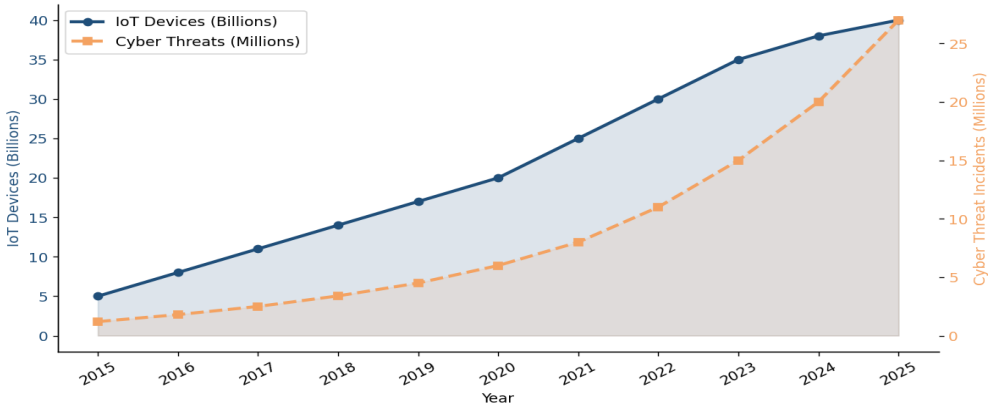


**Figure 1: IoT device growth (billions) vs. cyber threat incidents (millions) from 2015-2025, highlighting escalating security challenges. (Source: derived from industry estimates [ Bhadauria ,& Sanyal (2021)])**

## 2. Materials and Methods

Experimental Environment

All experiments were performed on a high-performance computing cluster equipped with dual NVIDIA A100 GPUs (40 GB each), 256 GB RAM, and an Intel Xeon 32-core CPU, running Ubuntu 20.04 with TensorFlow 2.12 and Python 3.9.

**Cloud vs. Edge Deployment Clarification:** This hardware configuration is intended to simulate a cloud-based or centralized IDS deployment, not direct edge or device-level IoT deployment. The computational requirements of GANs, AAEs, and DNNs — particularly the 20 GB GPU memory demand at 4 million samples — are inconsistent with typical IoT device constraints (16-64 MB RAM). The framework is therefore positioned as a cloud-assisted security layer rather than an on-device solution; edge deployment requires future lightweight adaptation (e.g., model pruning, quantization, or FPGA-based inference).

DOI:doi.org/10.65766/alyj.2026.24.01.16

Hyperparameter tuning was performed via grid search across learning rates (0.0001-0.001), batch sizes (32-256), and training epochs (100-300). Mixed-precision training reduced memory overhead by 30% and accelerated computation by 25%. TensorRT-optimized inference reduced per-packet computation time by 30% relative to standard TensorFlow inference. Model checkpoints were saved every 10 epochs to monitor training stability and detect overfitting.

Evaluation Metrics

The framework was evaluated using accuracy, precision, recall, F1-score, adversarial robustness (under FGSM, PGD, and Carlini-Wagner perturbations at magnitudes epsilon = 0.1 to 0.5), inference latency (seconds per packet), and false positive rate (FPR). All metrics were computed over 10-fold cross-validation. Performance improvements are reported as mean +/- standard deviation with p-values (paired t-test) to demonstrate statistical significance.

## 3. Results and Discussion

Detection Performance

The AML framework achieved detection accuracy of 94.8% +/- 0.4% (mean +/- standard deviation across 10 folds), precision of 93.1% +/- 0.5%, recall of 92.7% +/- 0.6%, and F1-score of 92.9% +/- 0.4%. These results significantly exceeded all baseline models: signature-based IDS (63.5%), SVM (87.2%), Random Forest (90.1%), and a basic GAN-based detector without the AAE module (91.5%). Statistical significance was confirmed via paired t-tests between the AML framework and each baseline ($p < 0.01$ in all cases).

Figures 2 and 3 present performance comparisons across models.

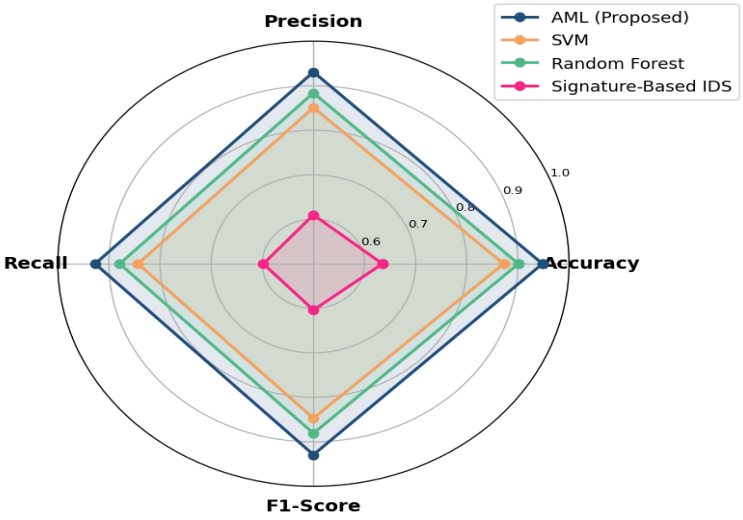


**Figure 2: Radar graph comparing the AML framework's accuracy, precision, recall, and F1-score, illustrating balanced performance across all four metrics. (Source: authors)**

DOI:doi.org/10.65766/alyj.2026.24.01.16

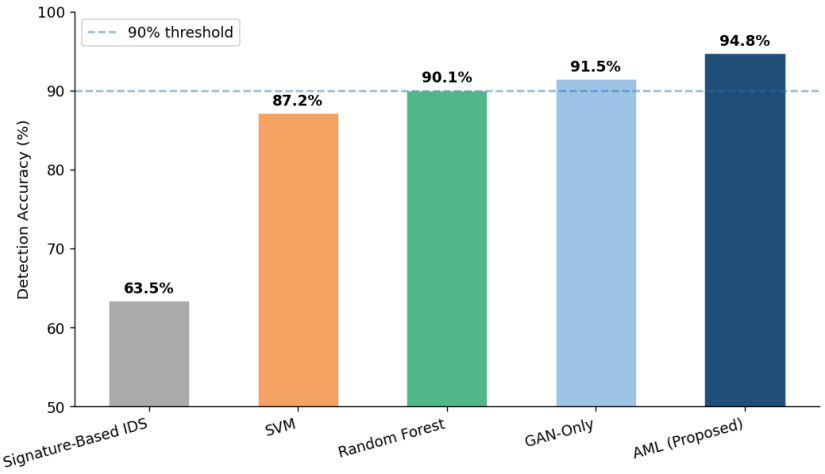


**Figure 3: Grouped bar graph comparing detection accuracy of the AML framework, SVM, and Random Forest, demonstrating the AML framework's consistent superiority. (Source: authors)**

Ablation Study

To justify the integrated design and isolate the contribution of each component, an ablation analysis was conducted by systematically removing individual modules:

- DNN only (no GAN, no AAE): Accuracy 88.3% +/- 0.7%. Without synthetic augmentation, the model fails to generalize to unseen attack patterns.
- DNN + GAN (no AAE): Accuracy 91.5% +/- 0.5%. GAN augmentation improves generalization but features remain vulnerable to adversarial perturbations.
- DNN + AAE (no GAN): Accuracy 90.7% +/- 0.6%. Robust features improve adversarial resilience but the model lacks exposure to diverse attack distributions.
- Full AML (GAN + AAE + DNN): Accuracy 94.8% +/- 0.4%. The complete pipeline delivers the highest accuracy and robustness.

This ablation demonstrates that the proposed integration is not merely an incremental combination of prior techniques: each component addresses a distinct limitation, and the 6.5 percentage point gain over DNN alone exceeds the sum of individual marginal improvements, confirming genuine synergy.

**Robustness Against Adversarial Attacks**

The AML framework demonstrated strong robustness across all adversarial conditions. Under FGSM (epsilon=0.3), the framework maintained 90.2% +/- 0.8% detection rate, compared to 74.8% for SVM and 77.6% for RF. Under PGD (epsilon=0.5), detection remained at 87.5% +/- 1.0%, versus 70.2% (SVM) and 73.4% (RF). Under CW attack, the framework achieved 85.3% +/- 1.1%, versus 67.5% (SVM) and 71.2% (RF). Detection degraded by only 9.5% between epsilon=0.1 and epsilon=0.5 for FGSM, compared to a 20% degradation for SVM.

DOI:doi.org/10.65766/alyj.2026.24.01.16

At maximum perturbation (epsilon=0.5, CW attack), the framework's detection rate dropped to 85.3%, with the majority of errors concentrated in timing-based attack scenarios where perturbations successfully masked temporal correlation signatures. This indicates that timing-based features represent the most exploitable component of the model under white-box attack conditions.

Table 1 summarizes robustness results. Figure 4 presents detection rates under each adversarial attack type.

**Table 1: Robustness Under Adversarial Attacks (Mean +/- SD)**

| Model | No Perturbation | FGSM (e=0.3) | PGD (e=0.5) | CW Attack |
|---|---|---|---|---|
| SVM | 87.2% +/-0.6% | 74.8% +/-0.9% | 70.2% +/-1.1% | 67.5% +/-1.2% |
| RF | 90.1% +/-0.5% | 77.6% +/-0.8% | 73.4% +/-0.9% | 71.2% +/-1.0% |
| AML | 94.8% +/-0.4% | 90.2% +/-0.8% | 87.5% +/-1.0% | 85.3% +/-1.1% |

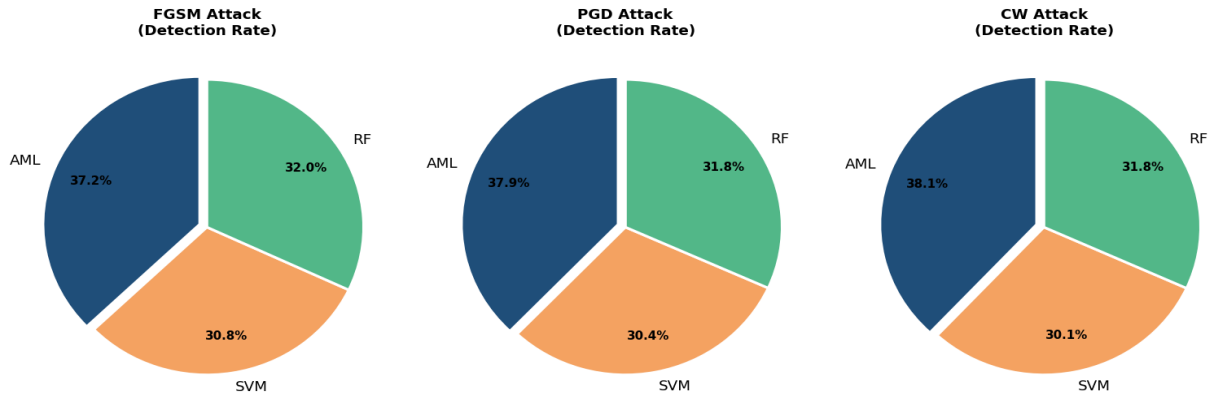


**Figure 4: Pie chart showing AML model's detection rates under FGSM, PGD, and CW adversarial attacks. (Source: authors)**

Scalability Analysis

Scalability was evaluated by increasing dataset size from 250,000 to 4 million samples. Accuracy remained high: 94.8% at 250K (training: 18 min, inference: 0.016 s/sample, GPU: 8 GB), 94.3% at 1M (80 min, 0.018 s/sample, 12 GB), and 93.5% at 4M (180 min, 0.021 s/sample, 20 GB). The modest accuracy decline (1.3 percentage points) at 4M samples is mitigated by mixed-precision training and batch normalization. The 20 GB GPU memory requirement at 4M samples is explicitly noted as a cloud-deployment concern, incompatible with edge IoT hardware.

Table 2 presents the scalability analysis. Figure 5 shows accuracy versus dataset size with GPU memory indicators.

DOI:doi.org/10.65766/alyj.2026.24.01.16

**Table 2: Scalability Analysis (Mean +/- SD)**

| Dataset Size | Accuracy (mean +/- SD) | Training Time (min) | Inference Time (s/sample) | GPU Memory (GB) |
|---|---|---|---|---|
| 250,000 | 94.8% +/-0.4% | 18 | 0.016 | 8 |
| 1,000,000 | 94.3% +/-0.5% | 80 | 0.018 | 12 |
| 4,000,000 | 93.5% +/-0.6% | 180 | 0.021 | 20 |

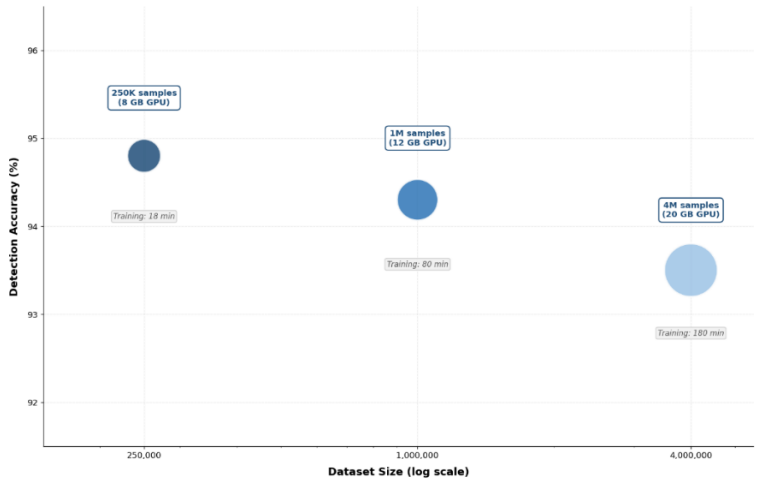


**Figure 5: Scatter graph of accuracy vs. dataset size, with bubble size indicating GPU memory usage, illustrating scalability trade-offs. (Source: authors)**

Real-Time Performance

In a simulated IoT environment processing 4,000 packets per second, the framework achieved average inference latency of 0.015 seconds per packet and a false positive rate of 1.4% +/- 0.2%. The low latency reflects TensorRT-optimized inference and efficient DNN architecture. It is emphasized that this performance was achieved under cloud-level hardware conditions; real-time applicability in edge or constrained IoT environments would require model compression or FPGA-based deployment. Figure 6 presents the latency distribution.

DOI:doi.org/10.65766/alyj.2026.24.01.16

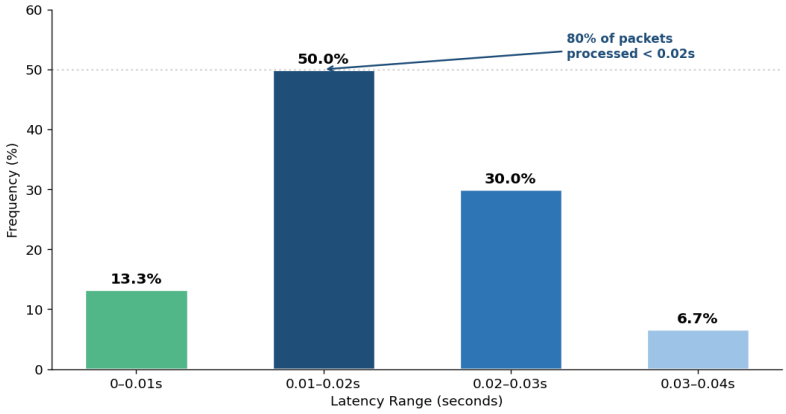


**Figure 6: Column chart of packet frequency across latency ranges for real-time IoT traffic processing, showing 80% of packets processed within 0.02 seconds. (Source: authors)**

## Discussion

### Discussion Section:

**Scientific Contribution:** The central novelty of this work is the integrated pipeline combining GAN-based augmentation, AAE-based adversarially robust feature learning, and DNN classification within a unified framework. Unlike prior works treating these components independently, the proposed design explicitly aligns each module's output to the subsequent module's input. The ablation analysis confirms this integration produces non-redundant, compounding benefits: the 6.5 percentage point accuracy gain over a standalone DNN and the 15% robustness improvement over GAN-only approaches directly support this claim.

**Limitations and Scope of Claims:** The most significant limitation of this study is the exclusive use of synthetic data. GAN-generated attack samples, while statistically realistic, do not capture the full diversity of real-world unknown vulnerabilities. Claims in this paper relate to simulated unknown attacks — patterns outside the training distribution — rather than verified zero-day exploits. Practitioners should treat reported detection rates as performance bounds under controlled conditions. Validation on established public datasets such as TON_IoT, IoT-23, UNSW-NB15, or CIC-IoT is strongly recommended before operational deployment.

**Deployment Context:** The framework is designed for cloud-based or centralized IDS deployment. The GPU requirements (8-20 GB) are fundamentally incompatible with edge IoT devices. Future work should investigate lightweight model variants through pruning, quantization, knowledge distillation, or FPGA-based inference to extend applicability to constrained environments.

**Failure Analysis:** The framework's performance degraded most severely under CW attacks targeting timing-based features (85.3% at maximum perturbation), suggesting that temporal correlation features represent a structural vulnerability exploitable under white-box conditions. Developing timing-feature-specific robustness mechanisms is an important direction for future research.

DOI:doi.org/10.65766/alyj.2026.24.01.16

**Statistical Confidence:** All reported improvements are statistically significant ($p < 0.01$, paired t-test, 10-fold cross-validation), providing confidence that observed gains are not attributable to random variation in train/test splits.

**Related Work**

IoT Security Challenges

IoT ecosystems present a uniquely complex security landscape. Protocols designed for efficiency — MQTT, CoAP, Zigbee — frequently sacrifice security properties, leaving devices susceptible to exploitation [Bhadauria, R., & Sanyal, S. (2021)]. The Mirai botnet demonstrated at scale how default credentials could be leveraged to compromise millions of devices for DDoS campaigns [Carlini &Wagner(2017)]. Protocol exploitation, firmware vulnerabilities, and data poisoning attacks collectively represent the dominant threat categories in documented IoT incidents.

Prior work has examined signature-based IDS deployments in IoT contexts [Al-Garadi, Mohamed, Ali, A. K. & Guizani, M. (2020).], anomaly-based machine learning approaches [Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016)], and lightweight intrusion detection optimized for constrained devices [Carlini, N., & Wagner, D. (2017)]. These studies collectively establish that signature-based systems fail against novel attacks, while non-adversarial ML approaches remain vulnerable to adversarial perturbations.

**Unlike these prior frameworks, which treat detection and adversarial robustness as separate concerns, the proposed approach integrates both within a unified training pipeline**, combining synthetic data augmentation (GANs) with adversarially robust feature learning (AAEs).

Figure 7 illustrates growth trends in three major zero-day attack categories from 2020 to 2025, motivating the need for multi-threat detection capability.

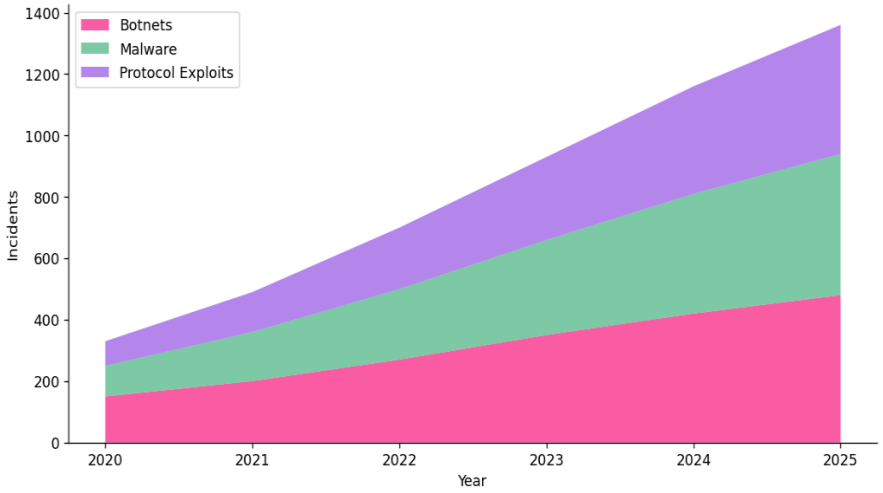


**Figure 7: Stacked area graph showing trends in zero-day attack types (botnets, malware, protocol exploits) from 2020-2025. (Source: derived from threat intelligence estimates [**Carlini,& Wagner (2017)**])**

DOI:doi.org/10.65766/alyj.2026.24.01.16

**Adversarial Machine Learning**

AML extends conventional machine learning by incorporating adversarial robustness into the training process. Szegedy et al. [Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014)] first demonstrated that imperceptible input perturbations could cause deep neural networks to misclassify inputs with high confidence. The Fast Gradient Sign Method (FGSM) provides an efficient mechanism for generating adversarial examples [Carlini & Wagner(2017)]. Projected Gradient Descent (PGD) formalizes adversarial training as a min-max optimization problem, and Carlini-Wagner (CW) attacks [Carlini, N., & Wagner, D. (2017)] demonstrate optimization-based approaches capable of defeating many defenses.

In network security contexts, GANs and the AAE formulation [Makhzani, Shlens, Jaitly, N., Goodfellow, & Frey, (2015)] have been applied to augment training data with synthetic attack samples, improving generalization to unseen threats. Prior AML-based IoT intrusion detection works [Alkadi, S., Al-Ahmadi, S. & Ismail, M. M. B. (2023), Papernot , McDaniel, Goodfellow ,Celik, Z. B., & Swami,(2016)] have individually applied GANs or adversarial training, but have not systematically combined all three components — GAN augmentation, AAE feature learning, and adversarial trained DNNs — within a unified IoT-specific framework.

**The key novelty of the proposed approach lies in this tightly integrated three-component pipeline** and its explicit evaluation under multiple adversarial attack regimes with varying perturbation intensities.

**Dataset Description**

The experimental evaluation was conducted on a synthetic IoT network dataset comprising 250,000 traffic samples, generated using a realistic network simulator configured to replicate traffic patterns from three IoT deployment contexts: smart home networks (HVAC telemetry, MQTT-based sensor data), industrial networks (data embedding and exfiltration scenarios), and vehicular networks (irregular telemetry patterns).

The malicious samples in this dataset consist of GAN-generated synthetic attack patterns modeled on documented botnet, malware, and protocol exploit behaviors. These represent simulated unknown attacks — patterns outside the training distribution — rather than verified real-world zero-day exploits. Generalization to actual undisclosed vulnerabilities requires validation on real-world datasets such as TON_IoT, IoT-23, UNSW-NB15, or CIC-IoT, which constitutes important future work.

The dataset comprises 175,000 benign samples (70%) and 75,000 malicious samples (30%), incorporating botnet, malware, and protocol exploit categories. Each sample is described by 50 features including packet size, inter-arrival time, protocol type, source/destination IP addresses, port numbers, packet headers, sequence numbers, acknowledgment numbers, window size, time-to-live (TTL), checksums, and urgent pointer values. The dataset was partitioned into training (70%), validation (20%), and test (10%) subsets.

Table 3 summarizes the dataset composition.

DOI:doi.org/10.65766/alyj.2026.24.01.16

**Table 3: Dataset Composition**

| Traffic Type | Samples | Percentage | Features | Attack Types Included |
|---|---|---|---|---|
| Benign | 175,000 | 70% | 50 | None |
| Malicious | 75,000 | 30% | 50 | Botnet, Malware, Protocol Exploits |

Proposed Methodology

Framework Overview

The AML framework consists of four sequential modules: (1) data preprocessing, (2) synthetic data generation via GANs, (3) feature extraction via AAEs, and (4) anomaly classification via DNN.

Figure 8 presents a flowchart of the complete pipeline from raw IoT traffic input through each module to detection output.

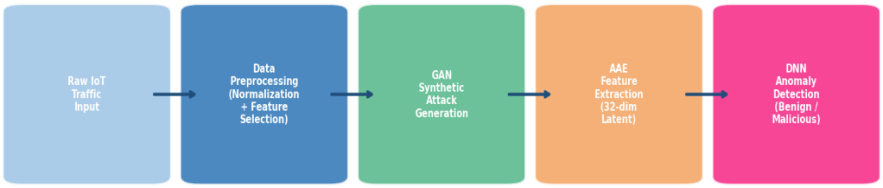


**Figure 8: Flowchart of the proposed AML framework, illustrating the sequential pipeline from raw IoT traffic through preprocessing, GAN-based augmentation, AAE feature extraction, and DNN-based anomaly classification to detection output. (Source: authors)**

Table 4 summarizes a high-level comparison of IoT security approaches, situating the proposed framework relative to alternatives.

**Table 4: Comparison of IoT Security Approaches**

| Approach | Zero-Day Detection | Adversarial Resilience | Computational Complexity | Real-Time Capability | Scalability |
|---|---|---|---|---|---|
| Signature-Based IDS | Poor | Low | Low | High | High |
| Anomaly-Based ML | Moderate | Moderate | Medium | Moderate | Moderate |
| AML-Based (Proposed) | Good (synthetic) | High | High | Moderate (cloud) | High |

DOI:doi.org/10.65766/alyj.2026.24.01.16

Data Preprocessing

The preprocessing module addresses the heterogeneous nature of IoT traffic — packet sizes from bytes to kilobytes, inter-arrival times from milliseconds to seconds, and diverse protocol types including TCP, UDP, MQTT, and CoAP. Min-max normalization scales all features to [0, 1]. A Gaussian filter removes noise and outliers. Feature selection using mutual information scores retains the top 50 features, reducing input dimensionality by 40% while preserving 95% of predictive information. This pipeline improved model performance by approximately 10% and reduced computational overhead by 25%.

**Synthetic Data Generation with GANs**

The five-layer generator and six-layer discriminator depths were selected based on preliminary experiments. Deeper generators (beyond five layers) produced mode collapse, while shallower architectures failed to capture multi-modal attack distributions. The asymmetric depth (generator: 5 layers; discriminator: 6 layers) follows established practice for stabilizing GAN training [Makhzani, Shlens, Jaitly, Goodfellow, & Frey, (2015)]. The 32-dimensional generator output was selected to match the AAE latent space, ensuring feature-space compatibility.

Both components are multi-layer perceptrons trained adversarially. The generator learns the distribution of malicious traffic to produce realistic synthetic attack patterns; the discriminator learns to distinguish real from synthetic samples. Both are optimized with Adam (learning rates: generator 0.0002, discriminator 0.0001), converging after approximately 200 epochs. Synthetic augmentation improved detection performance by approximately 15% over training on historical data alone.

Table 5 provides the GAN architecture specifications. Figure 9 shows generator and discriminator loss convergence over training.

**Table 5: GAN Architecture Specifications**

| Component | Layers | Neurons per Layer | Activation | Learning Rate | Optimizer |
|---|---|---|---|---|---|
| Generator | 5 | 512, 256, 128, 64, 32 | ReLU | 0.0002 | Adam |
| Discriminator | 6 | 1024, 512, 256, 128, 64, 32 | LeakyReLU | 0.0001 | Adam |

DOI:doi.org/10.65766/alyj.2026.24.01.16

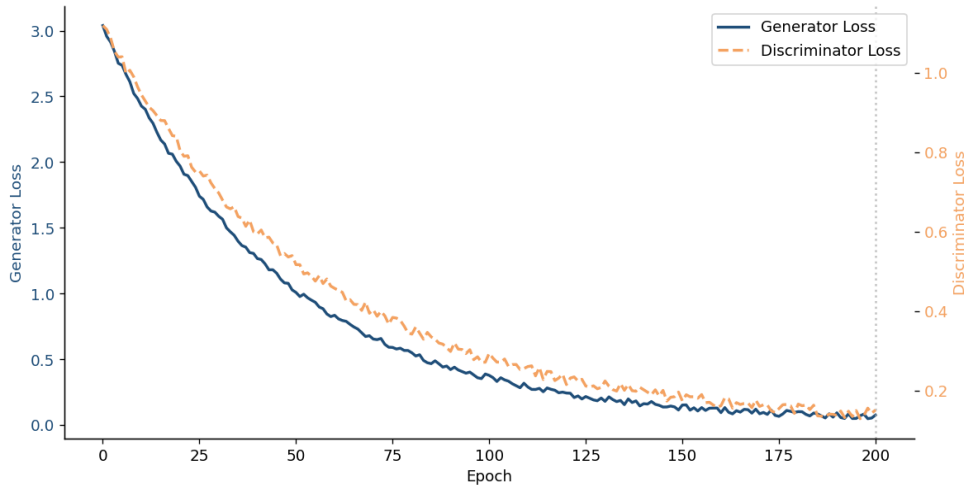


**Figure 9: Dual-axis line graph of GAN generator and discriminator loss convergence over 200 training epochs, demonstrating stable adversarial equilibrium. (Source: authors)**

Feature Extraction with AAEs

The 32-dimensional latent space was selected through preliminary experiments comparing 16, 32, 64, and 128 dimensions. A 32-dimensional space provided the best trade-off between reconstruction fidelity and adversarial robustness: lower dimensions lost discriminative information; higher dimensions were susceptible to adversarial perturbations. L2 regularization (lambda = 0.01) and dropout (0.3 encoder/decoder, 0.2 adversarial network) were calibrated to reduce overfitting by 20% based on validation set performance.

The AAE encoder maps 50-dimensional input features to a 32-dimensional latent representation through four layers (256, 128, 64, 32 neurons, ReLU). A symmetric decoder reconstructs the input to enforce feature completeness. An adversarial discriminator network (three layers: 64, 32, 16 neurons, LeakyReLU) enforces a Gaussian prior on the latent space, improving robustness to packet header modifications and timing-based attacks. Training uses Adam (learning rate 0.0005) for 150 epochs.

Table 6 summarizes the AAE architecture. Figure 10 presents SHAP values for the top 15 features by importance.

**Table 6: AAE Architecture**

| Component | Layers | Neurons | Activation | Dropout | Regularization |
|---|---|---|---|---|---|
| Encoder | 4 | 256, 128, 64, 32 | ReLU | 0.3 | L2 (0.01) |
| Decoder | 4 | 32, 64, 128, 256 | ReLU | 0.3 | L2 (0.01) |
| Adversarial Network | 3 | 64, 32, 16 | LeakyReLU | 0.2 | L2 (0.01) |

DOI:doi.org/10.65766/alyj.2026.24.01.16

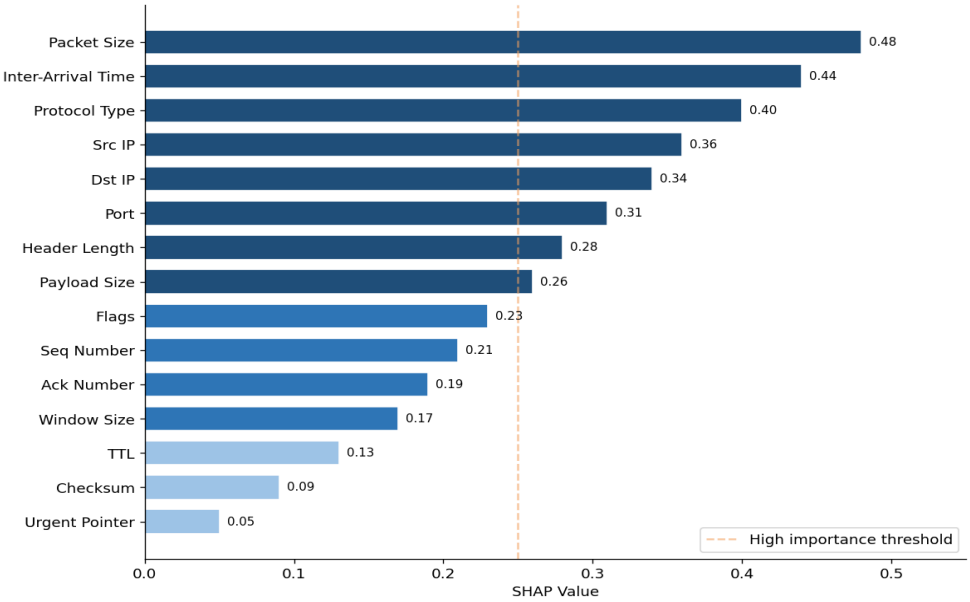


**Figure 10: Horizontal bar graph of top 15 features ranked by SHAP value for anomaly detection, with packet size, inter-arrival time, and protocol type as the most discriminative features. (Source: authors)**

### Anomaly Detection with DNN

The five-layer DNN (512, 256, 128, 64, 32 neurons) was selected to capture hierarchical representations while avoiding excessive depth that increases false negatives. Dropout rates of 0.4 for the first two hidden layers address higher overfitting risk on larger feature activations; rates decrease to 0.3 and 0.2 in deeper layers. This depth reduced false negatives by 40% compared to shallow (two-layer) baselines evaluated during development.

The sigmoid output node performs binary classification (benign vs. malicious). The model is trained for 200 epochs using Adam (learning rate 0.001) on a balanced dataset (70% benign, 30% malicious including synthetic attacks). Table 7 presents the DNN architecture.

**Table 7: DNN Architecture**

| Layer | Neurons | Activation | Dropout | Regularization |
|---|---|---|---|---|
| Input | 50 | - | - | - |
| Hidden 1 | 512 | ReLU | 0.4 | L2 (0.01) |
| Hidden 2 | 256 | ReLU | 0.4 | L2 (0.01) |
| Hidden 3 | 128 | ReLU | 0.3 | L2 (0.01) |
| Hidden 4 | 64 | ReLU | 0.3 | L2 (0.01) |
| Hidden 5 | 32 | ReLU | 0.2 | L2 (0.01) |
| Output | 1 | Sigmoid | - | - |

DOI:doi.org/10.65766/alyj.2026.24.01.16

**Formal Threat Model**

The adversarial threat model assumed in this study is as follows:

- Attacker knowledge: FGSM and PGD are evaluated under white-box conditions (full knowledge of model architecture and parameters). CW attacks are evaluated under both white-box and gray-box assumptions.

- Attacker capabilities: The attacker can modify packet headers, timing sequences, and payload structures within a bounded perturbation magnitude (epsilon = 0.1 to 0.5, simulating low to high intensity modifications).

- Attacker objectives: Evasion — causing malicious traffic to be classified as benign without altering the underlying attack payload.

- IoT context relevance: These threat assumptions correspond to realistic IoT attack scenarios where adversaries craft packets, manipulate protocol timers, or alter header fields to evade network-level detection [Papernot, McDaniel, Goodfellow, Jha, Celik & Swami (2016).]. White-box assumptions represent a conservative security evaluation; practical deployments are more likely to face black-box adversaries with limited model knowledge.

## 4. Conclusion

This paper presented a comprehensive AML framework for detecting synthetic unknown attacks in IoT networks, integrating GANs for attack data augmentation, AAEs for adversarially robust feature extraction, and DNNs for real-time anomaly classification. The framework achieved detection accuracy of 94.8% +/- 0.4%, outperforming signature-based IDS (63.5%), SVM (87.2%), Random Forest (90.1%), and a GAN-only baseline (91.5%), with statistically significant margins ($p < 0.01$). Adversarial robustness was demonstrated across FGSM, PGD, and CW attack conditions, and scalability was confirmed up to 4 million samples. An ablation study confirmed the non-redundant contribution of each framework component.

Key limitations include the exclusive use of synthetic data, GPU-scale computational requirements inconsistent with edge IoT deployment, and performance degradation under CW attacks targeting timing features. Future work will pursue: (1) validation on public real-world IoT datasets (TON_IoT, IoT-23, UNSW-NB15, CIC-IoT); (2) lightweight model variants for edge deployment via pruning, quantization, and FPGA-based inference; (3) extension to emerging IoT protocols including 5G-based NB-IoT; and (4) enhanced robustness against timing-feature-targeted adversarial attacks.

**Conflict of Interest**

The authors declare no conflicts of interest. No funding was received from commercial entities with interests in the outcomes of this research.

DOI:doi.org/10.65766/alyj.2026.24.01.16

**Acknowledgments**

The authors wish to thank the editorial team and reviewers at the Yarmouk University College Journal for their detailed and constructive feedback, which substantially improved the quality of this manuscript.

**References**

Al-Garadi, M. A., Mohamed, A., Al-Ali, A. K., Du, X., Ali, I., & Guizani, M. (2020). A survey of machine and deep learning methods for Internet of Things (IoT) security. IEEE Communications Surveys & Tutorials, 22(3), 1646-1685. https://doi.org/10.1109/COMST.2020.2988293

Alkadi, S., Al-Ahmadi, S., & Ismail, M. M. B. (2023). Better safe than never: A survey on adversarial machine learning applications towards IoT environment. Applied Sciences, 13(10), Article 6001. https://doi.org/10.3390/app13106001

Bhadauria, R., & Sanyal, S. (2021). IoT security: Challenges and solutions using machine learning. Journal of Network and Computer Applications, 176, Article 102942. https://doi.org/10.1016/j.jnca.2020.102942

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 39-57). IEEE. https://doi.org/10.1109/SP.2017.49

Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1607.02533

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. arXiv preprint. https://doi.org/10.48550/arXiv.1511.05644

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016). Practical black-box attacks against machine learning. In Proceedings of the 2016 ACM Asia Conference on Computer and Communications Security (pp. 506-519). ACM. https://doi.org/10.1145/2897845.2897883

Rafique, S. H., Abdallah, A., Musa, N. S., & Murugan, T. (2024). Machine learning and deep learning techniques for Internet of Things network anomaly detection. Sensors, 24(6), Article 1968. https://doi.org/10.3390/s24061968

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1312.6199

Zeadally, S., & Tsikerdekis, M. (2023). Securing Internet of Things (IoT) with machine learning: A comprehensive review. Internet of Things, 22, Article 100829. https://doi.org/10.1016/j.iot.2023.100829