NOUN Interdisciplinary Journal of Computing, E-learning & Application (NOUN-IJCEA)
https://ijcea.nou.edu.ng
Volume 1 Number 1, April 2026 pp. 13-29

Implementing an Intrusion Detection System in Internet of Things Resource-Constraint Environment Using Hybrid Extreme Gradient Boost and Gated Recurrent Unit

[*1]Abraham Eseoghene Evwiekpaefe, [1]Isah Rambo Saidu, [2]Ismail Yunus, [1]Darius Tienhua Chinyio

[1]Department of Computer Science, Nigerian Defence Academy, Kaduna Nigeria.
[2]Department of Computer Science, Alhikma Polytechnic Karu, Nasarawa State, Nigeria.
aeevwiekpaefe@nda.edu.ng, rambo@nda.edu.ng, yunus.ismail2022@nda.edu.ng, dtchinyo@nda.edu.ng
*Corresponding Author: aeevwiekpaefe@nda.edu.ng

ABSTRACT

Intrusion Detection Systems (IDS) for Internet of Things (IoT) environments increasingly rely on deep learning models to detect complex attack patterns. While recent architectures achieve high accuracy under benign conditions, their resilience against adversarial evasion attacks remains insufficiently explored. This study evaluated adversarial robustness as primary design objective rather than secondary performance attribute. A lightweight hybrid XGBoost–GRU model was proposed and compared against attention-based LSTM (AT-LSTM). Gradient-based evasion attacks that includes Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) were employed under bounded perturbation budgets to simulate realistic adversarial behaviour. Experimental results on the UNSW-NB15 dataset showed that with the FGSM attack, XGBoost–GRU model retains approximately 90% accuracy at ε = 0.10 whereas the AT-LSTM degrades more sharply to approximately 86%. Under the PGD attacks which represent stronger and more adaptive adversarial strategy, the AT-LSTM experienced severe performance collapse with adversarial accuracy dropping to approximately 43% at ε = 0.10 while the developed XGBoost–GRU model maintains adversarial accuracy above 70% under the same conditions. Also, for the Attack Success Rate (ASR) analysis, more than half of malicious samples successfully evade detection in the AT-LSTM under PGD attacks (58%) whereas the XGBoost–GRU model limits successful evasions to less than 30%. The results from the developed XGBoost–GRU model consistently maintained higher adversarial accuracy, lower attack success rates, and superior robust accuracy retention under increased adversarial evasion. The findings demonstrate that architectural simplicity can enhance detection stability under adversarial pressure for intrusion detection in IoT systems.

Keywords: Adversarial Machine Learning, Deep Learning, Internet of Things, Intrusion Detection Systems, Lightweight

1. INTRODUCTION

This decades have experienced rapid expansion of the Internet of Things (IoT) into contemporary digital infrastructures which enables large-scale connectivity among physical devices deployed in domains such as healthcare, smart cities, and industrial automation (Vitorino et. al., 2025). Though this pervasive connectivity supports real-time monitoring and automation, it also substantially

enlarges cyber-attack surface. Many IoT devices operate under strict resource constraints such as limited computational capacity, memory, and energy availability which restricts deployment of conventional security mechanisms (Liu et al., 2024). As a result,signature-based and rule-driven Intrusion Detection Systems (IDS) struggle to cope with evolving attack behaviors in IoT environments (More et al., 2024). To overcome these limitations, machine learning (ML) and deep learning (DL) approaches have been increasingly adopted to enhance detection capability and adaptability. Recurrent neural network (RNN) architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models have been widely applied due to their ability to capture temporal dependencies in network traffic (Ali et al., 2024; Aleesa et al., 2021; Alnuaimi and Albaldawi, 2024).

Several studies report strong detection performance on benchmark datasets using deep architectures. However, many of these models rely on complex multi-layer designs that impose substantial computational overhead. For example, the hierarchical attention-based LSTM (AT-LSTM) proposed by (Alsharaiah et al., 2024) achieves high classification accuracy on the UNSW-NB15 dataset, yet its architectural complexity raises concerns regarding inference latency and deployability on edge and IoT devices. In addition to efficiency constraints, a growing body of research has identified adversarial robustness as critical and underexplored challenge in IoT intrusion detection (Layeghy et al., 2023). Adversarial machine learning (AML) techniques enable attackers to craft carefully perturbed inputs that remain statistically similar to benign traffic while inducing misclassification. Such evasion strategies pose serious threat to IDS models deployed in operational environments (Asharf et al., 2020). High-capacity architectures like attention-enhanced LSTM networks may be particularly vulnerable due to their reliance on high-dimensional feature representations which can unintentionally amplify adversarial sensitivity (Armijos and Cuenca, 2023). On the other hand, hybrid IDS frameworks that incorporate feature selection mechanisms such as eXtreme Gradient Boosting (XGBoost) offer the potential to limit adversarial influence by reducing input dimensionality and suppressing irrelevant or noisy features (Alharthi et al., 2025).

This study extends prior work on lightweight intrusion detection by shifting the focus from computational efficiency to adversarial resilience. Using the previously established XGBoost–GRU architecture as baseline, the paper conducts a systematic security stress test to evaluate detection stability under adversarial evasion attacks. The proposed model is assessed against the attention-based LSTM framework of (Alsharaiah et al., 2024) on the UNSW-NB15 dataset with emphasis on performance degradation, robustness under attack, and classification consistency in hostile conditions.

2. LITERATURE REVIEW

Research on intrusion detection in Internet of Things (IoT) environments has intensified due to the increasing exposure of resource-constrained devices to sophisticated cyber threats (Alanzi and Aljuhani, 2023). Existing studies broadly span deep learning–based intrusion detection, hybrid machine learning frameworks and adversarial machine learning applied to security systems. Despite progress across these areas, their integration remains limited, particularly with respect to adversarial robustness in lightweight IoT intrusion detection systems.

Deep learning approaches have been widely adopted for intrusion detection because of their capacity to learn complex and nonlinear relationships in network traffic (Sarker, 2021). Recurrent neural network architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been extensively used to capture temporal dependencies associated with multi-stage attacks and evolving intrusion patterns (Nosouhian et al., 2021). Several studies report high detection accuracy on benchmark datasets such as UNSW-NB15 and NSL-KDD using LSTM-based models (Zhang et al., 2022). Recent work has incorporated attention mechanisms to enhance temporal focus and feature relevance. Alsharaiah (2024) proposed hierarchical attention-based LSTM (AT-LSTM) model that

achieved strong detection performance on the UNSW-NB15 dataset. While attention enhanced architectures improve representational capacity, their multi-layer design substantially increases computational cost and inference latency, which limits feasibility for real-time deployment on IoT and edge devices (Alwahedi et al., 2024). Moreover, deep architectures optimized primarily for accuracy often exhibit high sensitivity to small perturbations in input space. Studies in adversarial machine learning suggest that complex decision boundaries and high-dimensional feature representations can increase susceptibility to evasion attacks, an issue that remains under-examined in conventional IDS evaluations (Mienye and Swart, 2024).

To address efficiency limitations of deep learning models, hybrid intrusion detection systems combining machine learning and deep learning techniques have gained attention (Halbouni et al., 2022). Tree-based ensemble methods that includes decision trees, random forests, and gradient boosting, are frequently employed for feature selection and dimensionality reduction (Soon et al., 2024). Among these algorithms, eXtreme Gradient Boosting (XGBoost) has demonstrated strong performance in handling high-dimensional and imbalanced intrusion datasets (Chen and Guestrin, 2024). Several studies integrate XGBoost-based feature selection with deep learning classifiers to improve training efficiency and inference speed while maintaining competitive accuracy (Alraba'nah et al., 2025). Although such hybrid frameworks reduce redundancy and computational overhead, they are commonly evaluated under non-adversarial conditions with emphasis placed on accuracy, precision, and recall rather than robustness to malicious input manipulation (Rahman et al., 2024). Importantly, feature selection in existing IDS research is typically framed as optimization strategy rather than security-oriented design choice (Soon et al., 2024). The potential of reduced feature spaces to limit adversarial exploitability has received minimal empirical investigation in IoT intrusion detection research.

Adversarial machine learning has now regarded as critical concern for security-sensitive ML applications such as intrusion detection systems (Ahmed et al., 2023). Adversarial evasion attacks involve introducing carefully designed perturbations into input features to induce misclassification while preserving semantic validity. Prior studies demonstrate that both traditional machine learning models and deep neural networks can experience significant performance degradation when subjected to such attacks (Dash et al., 2024). Existing adversarial IDS research primarily focuses on attack generation techniques or defense mechanisms such as adversarial training, ensemble learning, or input reconstruction. While these methods improve resilience, they often introduce additional training complexity and runtime overhead which limit their suitability for deployment in resource-constrained IoT environments (Khaw et al., 2021). Furthermore, most adversarial evaluations consider individual models in isolation rather than conducting controlled comparisons between architectures of differing complexity (Meena and Indian, 2025). As a result, the relationship between model design, feature dimensionality, and adversarial vulnerability remains insufficiently understood.

2.1 Research Gap

Although advances have been made in deep learning-based IDS hybrid intrusion detection and adversarial machine learning, their convergence remains limited. IoT intrusion detection research continues to prioritize detection accuracy and efficiency while adversarial studies emphasize attack effectiveness or defense augmentation, often overlooking deployment (Alwahedi et al., 2024).

This study addresses this gap by examining adversarial robustness as architectural property rather than externally imposed defense. The work provides empirical insight into how architectural simplicity influence detection stability by developing a lightweight XGBoost–GRU hybrid model and subsequently comparing the developed lightweight XGBoost–GRU hybrid model with attention-based LSTM framework under identical adversarial conditions on the UNSW-NB15 dataset (Moustafa and

Slay, 2015). This perspective supports the design of an intrusion detection systems that balance robustness, efficiency, and practical deployability in IoT environments.

3. METHODOLOGY

This study adopts structured adversarial evaluation framework designed to assess detection stability under evasion attacks rather than peak performance under benign conditions.

Figure 1 presents the structured adversarial evaluation framework adopted in this study. First, both models are trained on clean data to establish baseline performance. Second, adversarial samples are generated using gradient-based attacks FGSM and PGD under bounded perturbation budgets. Finally, robustness is evaluated using adversarial accuracy, attack success rate, and resilience factor. This framework ensures a fair and reproducible comparison between lightweight and complex architectures.

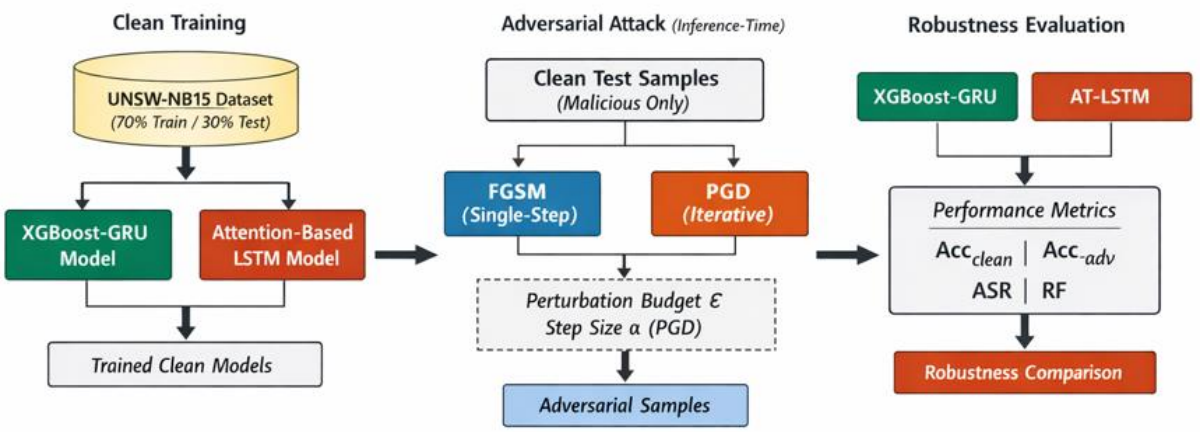

Figure 1: Overview of the proposed adversarial evaluation framework

First, victim models are trained on clean unperturbed datasets to establish baseline intrusion detection performance. Data preprocessing was conducted systematically to improve data quality and learning efficiency. The processes include data cleaning, handling of missing values, feature encoding, normalization, and dataset partitioning. Duplicate records and inconsistent entries were removed and numerical outliers were examined to minimize distortion during training. Missing values were addressed using statistical imputation or removed when proportion of missing data was excessive to conform with established practices in intrusion detection research (kasongo and Sun, 2020). Categorical features such as protocol type and service were transformed using one-hot encoding, while target labels were encoded into numerical form. All numerical features were normalized to ensure consistent scaling across datasets. Second, adversarial samples are generated at inference time using gradient-based evasion techniques under bounded perturbation budget. Finally, model robustness is evaluated by measuring performance degradation and evasion success under increasing adversarial pressure. All experiments are conducted using identical data splits of 70% for training and 30% for testing and validation, attack configurations, and perturbation constraints across models. This ensures that observed differences in robustness arise from architectural design choices rather than experimental bias.

3.1 Victim Model Architectures

To examine the relationship between architectural complexity and adversarial resilience, two contrasting intrusion detection models are evaluated. Figure 2 contrasts the lightweight hybrid XGBoost–GRU architecture with the high complexity attention based LSTM. The lightweight model reduces feature dimensionality using XGBoost before applying temporal learning with GRU. In

contrast, the AT-LSTM processes the full feature space and applies attention over temporal states. This architectural difference enables us to examine whether reduced complexity improves adversarial resilience.

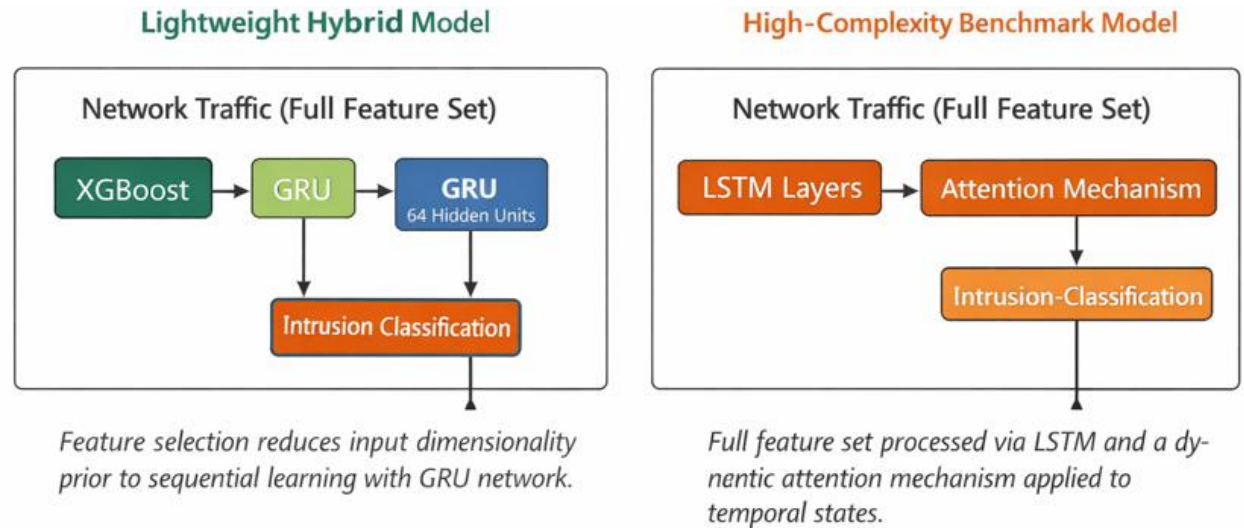

Figure 2: Comparison of evaluated Model Architectures

3.2 Lightweight Hybrid Model: XGBoost–GRU

The proposed model follows hybrid design that combines feature selection with sequential learning. In the first stage, eXtreme Gradient Boosting (XGBoost) is employed as a feature relevance filter to reduce original high-dimensional feature space to compact subset of the most informative attributes. This reduction aims to eliminate redundant and low-utility features prior to sequence modeling. In the second stage, a Gated Recurrent Unit (GRU) network with 64 hidden units captures temporal dependencies in network traffic. Compared to deeper recurrent architectures, the GRU offers simplified gating structure that reduces model complexity and limits sensitivity to minor input perturbations. This design prioritizes stability and efficiency while preserving temporal learning capability.

3.3 High-Complexity Benchmark Model: Attention-Based LSTM

As a representative high-capacity deep learning architecture, this study reproduces the hierarchical attention-based LSTM (AT-LSTM) model proposed by (Alsharaiah et al., 2024). The model processes full feature set without dimensionality reduction and employs attention mechanism to weight temporal states dynamically. While attention-based architectures have demonstrated strong accuracy on clean intrusion datasets, their reliance on dense feature representations and multi-stage transformations may increase sensitivity to gradient-based manipulation. This model serves as benchmark for evaluating whether architectural simplicity contributes to improved detection stability under adversarial conditions.

3.4 Threat Model and Adversarial Capabilities

The evaluation assumes constrained but adversarial strong threat model consistent with established adversarial machine learning practice (Mienye and Swart, 2024).

Adversary knowledge

A white-box setting is assumed where the attacker has full access to model parameters, architecture, and gradients. This represents worst-case evaluation scenario and provides conservative estimates of robustness.

Attack objective

The adversary seeks to perform evasion attacks at inference time by causing malicious traffic instances to be misclassified as benign. Attacks are applied exclusively to malicious samples to reflect realistic evasion objectives.

Constraints

Perturbations are restricted to numerical traffic features and bounded by maximum perturbation budget ε. Feature modifications must preserve functional validity and protocol compliance. The adversary has no ability to influence training data, feature selection procedures, or model parameters. Data poisoning, backdoor insertion, and retraining attacks are excluded from scope.

3.5 Adversarial Sample Generation

To simulate realistic evasion strategies, this study employs two widely used gradient-based adversarial attacks that differ in strength and computational cost (Nosouhian et al., 2021), Figure 3 illustrates the adversarial perturbation strategies used in this study. FGSM applies a single-step gradient-based perturbation, representing a fast and computationally inexpensive attack. PGD extends this into an iterative process, allowing the adversary to refine perturbations within an ε-bounded region.

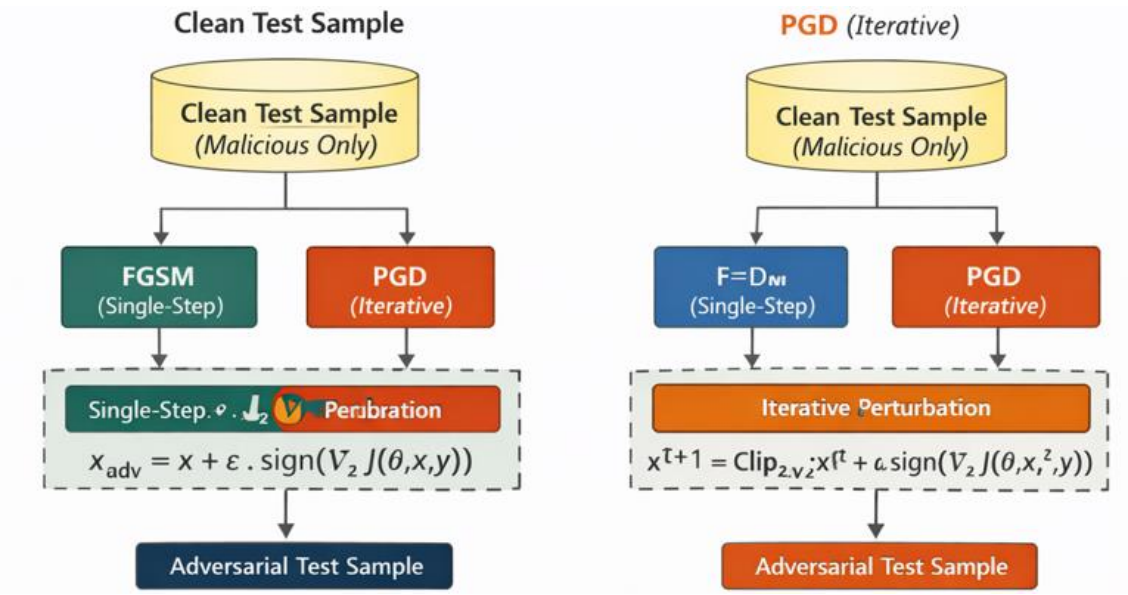

Figure 3: Comparison of adversarial input perturbation strategies

3.6 Fast Gradient Sign Method (FGSM)

FGSM is single-step attack that perturbs input samples in the direction of gradient that maximizes the model loss (Rahman et al., 2025; Mienye and Swart, 2024). Adversarial samples are generated as:

$$x_{adv} = x + \epsilon \cdot sign\left(\nabla_z J(\theta, x, y)\right) \qquad (1)$$

Where x denotes original input feature vector representing network traffic instance; x_{adv} is the adversarially perturbed version of the input; θ represents trained model parameters; y denotes true class label of input sample; $J(\theta, x, y)$ is loss function used during model training (e.g., cross-entropy); $\nabla_z J(\theta, x, y)$ is the gradient of loss with respect to the input features; *sign(·)* extracts sign of each gradient component to produce direction of maximal loss increase; and ϵ is perturbation magnitude that controls the strength of the attack (Nosouhian et al., 2021). The perturbation is applied uniformly across features in the direction that most rapidly increases classification error. Multiple values of ϵ are

evaluated to simulate varying attack intensities, ranging from mild feature manipulation to stronger evasion attempts. FGSM is computationally efficient and reflects fast low-cost attack strategy suitable for time-constrained adversaries (Alharthi and Djenori, 2025).

3.7 Projected Gradient Descent (PGD)

PGD extends FGSM into iterative optimization process to enable adversary to refine perturbations over multiple steps while enforcing strict constraint on total perturbation magnitude (Dash et al., 2024). Starting from original input *x*, PGD updates adversarial samples as:

$$x^{t+1} = Clip_{x,\epsilon}\left(x^t + \alpha \,.\, sign\left(\nabla_z J(\theta, x^t, y)\right)\right) \qquad (2)$$

Where x^t represents adversarial sample at iteration *t*; α is the step size that controls magnitude of each incremental update; $\nabla_z J(\theta, x^t, y)$ denotes gradient of the loss with respect to current adversarial input; and $Clip_{x,\epsilon}(.)$ projects perturbed sample back into an ϵ-bounded region centered around original input x to ensure that total perturbation remains within allowable limits (Meena and Choudhary, 2017). The clipping operation enforces constraint $||\, x^{t+1} - x|| \infty \leq \epsilon$, in or order to prevents excessive feature distortion that could invalidate network traffic semantics. PGD is widely regarded as strong first-order adversarial attack and is commonly used as a worst-case robustness benchmark due to its ability to iteratively exploit local model vulnerabilities (kasongo, 2023). Both FGSM and PGD capture rapid single-step evasion and adaptive multi-step adversarial behavior without introducing unrealistic assumptions regarding attacker capabilities or data manipulation (Aleesa et al., 2021).

3.8 Robustness Evaluation Metrics

To assess detection stability under adversarial conditions, this study extends conventional intrusion detection metrics by incorporating robustness-focused measures that explicitly quantify performance degradation and evasion effectiveness.

Clean Accuracy (Acc_{clean})

Clean Accuracy represents classification accuracy of the model when evaluated on original unmodified test dataset. This metric establishes a baseline level of detection capability under non-adversarial conditions and serves as reference point for measuring robustness loss.

Adversarial Accuracy (Acc_{adv})

Adversarial Accuracy measures model's classification accuracy when evaluated on adversarially perturbed test samples. It reflects the model's ability to maintain correct predictions in the presence of malicious feature manipulation. Higher values indicate stronger resistance to evasion attacks (Vitorino et al., 2025).

Attack Success Rate (ASR)

Attack Success Rate quantifies effectiveness of adversarial evasion strategies by measuring proportion of malicious samples that are incorrectly classified as benign under attack. It can be mathematically represented as follows:

$$ASR = \frac{Number\ of\ successful\ evasions}{Total\ number\ of\ malicious\ samples}\ 100\% \qquad (3)$$

A higher ASR indicates greater vulnerability as larger fraction of attacks successfully bypass intrusion detection system (Liu et al., 2024).

Resilience Factor (RF)

To evaluate stability under worst-case adversarial pressure, Resilience Factor (RF) is defined as:

$$RF = \frac{Acc_{adv}(\epsilon_{max})}{Acc_{clean}} \quad (4)$$

where ϵ_{max}= 0.1 represents maximum perturbation magnitude that preserves feature plausibility and network traffic validity. RF measures proportion of baseline accuracy retained under strongest evaluated attack. Values closer to 1 indicate high stability, whereas lower values reflect significant performance collapse. This metric enables direct and normalized comparison between models with differing baseline accuracies and architectural complexity (Akif et al., 2025).

3.9 Experimental Environment and Implementation Details

All experiments were conducted on local workstation to ensure reproducibility and controlled evaluation conditions. The hardware environment consisted of an Intel Core i7 processor, 16 GB of RAM, and GPU with 6 GB VRAM. Where GPU acceleration was unavailable, models were executed in CPU mode to ensure consistency across experimental runs. The software environment was based on Ubuntu 20.04 LTS with Python 3.9 used as the primary programming language. The XGBoost implementation was developed using the Scikit-learn library, while deep learning models were implemented using TensorFlow/Keras. Adversarial attacks were generated using standard gradient-based procedures implemented within TensorFlow framework, following established adversarial evaluation practices (Le et al., 2022). All experiments were executed under identical preprocessing, training, and evaluation configurations to ensure fair comparison between victim models. Random seeds were fixed across all runs to reduce stochastic variation and improve result stability.

3.10 Algorithms for the Methodology Implementations

Algorithm 1: Training–Attack–Evaluation Pipeline

1: Split dataset D into training set D_train and test set D_test
2: Train victim model M using D_train
3: Evaluate M on clean test samples → Acc_clean
4: Initialize counter for successful evasions = 0
5: For each malicious sample x in D_test:
6: Generate adversarial sample x_adv using attack A with budget ε
7: Predict class of x_adv using model M
8: If x_adv is misclassified as benign:
9: Increment successful evasion counter
10: Evaluate M on full adversarial test set → Acc_adv
11: Compute ASR using successful evasion counter
12: Compute RF = Acc_adv / Acc_clean
13: Return Acc_clean, Acc_adv, ASR, RF

Algorithm 2: Lightweight XGBoost–GRU Model Construction

1: Train XGBoost classifier on X and Y
2: Compute feature importance scores
3: Select top k features to form reduced dataset X_reduced
4: Construct GRU network with fixed hidden units
5: Generate sequential samples from X_reduced
6: Train GRU network using sequential training data
7: Output trained XGBoost–GRU model

Algorithm 3: Attention-Based LSTM (AT-LSTM) Model Construction

1: Construct multi-layer LSTM network
2: Attach attention mechanism over temporal hidden states
3: Use full feature set without dimensionality reduction
4: Train network using sequential input samples
5: Learn attention weights jointly with model parameters
6: Output trained AT-LSTM model

Algorithm 4: FGSM Adversarial Sample Generation

1: Compute loss gradient with respect to input x
2: Compute signed gradient direction
3: Generate adversarial sample:
$x_{adv} = x + \varepsilon \times \text{sign}(\text{gradient})$
4: Return x_{adv}

Algorithm 5: PGD Adversarial Sample Generation

1: Initialize $x_{adv} = x$
2: For $t = 1$ to T:
3: Compute loss gradient with respect to x_{adv}
4: Update x_{adv} using step size α
5: Project x_{adv} into ε-bounded region around x
6: Return x_{adv}

4. RESULTS AND DISCUSSION

This section presents the empirical evaluation of adversarial resilience for the proposed XGBoost–GRU model and the benchmark attention-based LSTM (AT-LSTM). The analysis focuses on detection stability under evasion attacks rather than peak accuracy under benign conditions. Results are reported using clean accuracy, adversarial accuracy, attack success rate, and robustness-oriented metrics introduced in Section 3. All evaluations follow the controlled training attack evaluation pipeline to ensure fair comparison.

4.1 Baseline Detection Performance on Clean Data

Baseline performance is first evaluated using the clean unperturbed test set to establish reference detection capability. Under non adversarial conditions, the AT-LSTM achieves slightly higher classification accuracy of 95% than the lightweight XGBoost–GRU model's 93% as shown in Figure 4. This result is consistent with prior studies reporting strong performance for attention-based architectures on standard intrusion detection benchmarks.

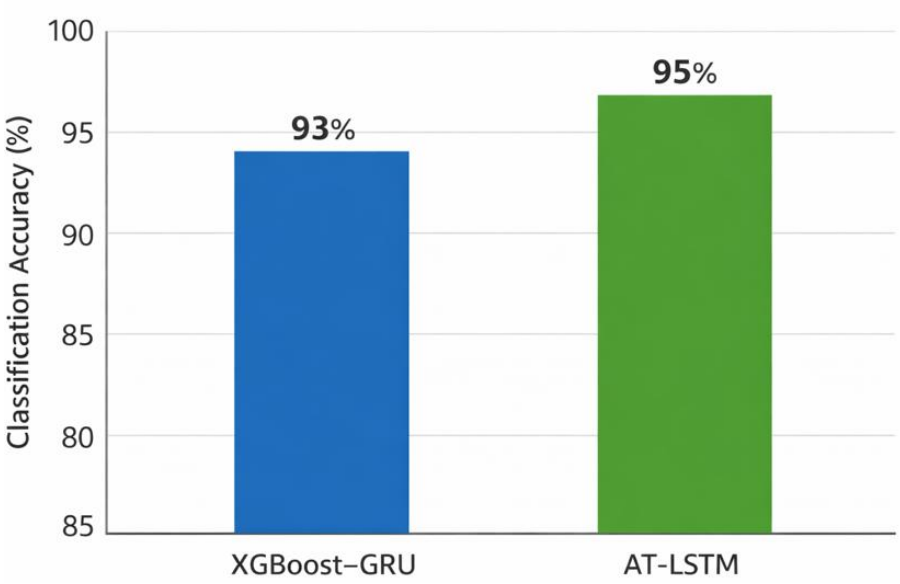

Figure 4: Baseline Detection Performance on Clean Data of the 2 Models

However, as illustrated later in subsequent results, superior clean accuracy does not necessarily translate into resilience under adversarial conditions. Baseline results therefore serve solely as a reference point for subsequent robustness evaluation.

4.2 Robustness Under FGSM-Based Evasion Attacks

Figure 5 illustrates adversarial accuracy degradation from 90% for the proposed XGBoost-GRU and AT-LSTM's 86% under Fast Gradient Sign Method (FGSM) attacks across increasing perturbation magnitudes. Both models experience performance decline as ε increases, confirming that even small feature perturbations can negatively impact intrusion detection accuracy.

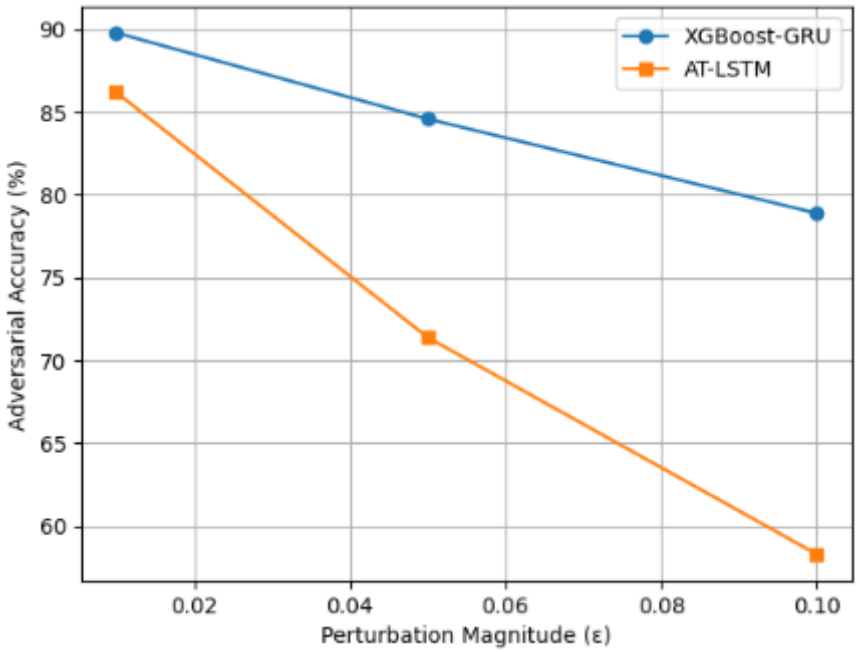

Figure 5: FGSM Attack Accuracy Degradation of the 2 models at ε = 0.10

The XGBoost–GRU model demonstrates a more gradual reduction in accuracy compared to the AT-LSTM. While the AT-LSTM exhibits a sharp performance drop at moderate perturbation levels, the

lightweight model retains a higher proportion of its baseline accuracy. This behaviour suggests that feature reduction and simplified recurrent structure limit the effectiveness of rapid, single-step adversarial manipulation. FGSM results highlight that lightweight architectures may offer improved resistance to fast, low-cost evasion strategies commonly associated with resource-constrained attackers.

4.3 Robustness Under PGD-Based Iterative Attacks

Projected Gradient Descent (PGD) attacks impose a stronger adversarial threat through iterative refinement of perturbations. Figure 6 presents adversarial accuracy trends under PGD across increasing ε values. Under iterative attack conditions, the AT-LSTM experiences substantial accuracy degradation from 80% down to about 43% with performance collapsing at higher perturbation levels. In contrast, the XGBoost–GRU model maintains noticeably higher accuracy across all evaluated ε values, 90% to about 73%. The widening performance gap under PGD indicates that complex attention mechanisms may expose richer gradient information that adversaries can exploit over multiple iterations.

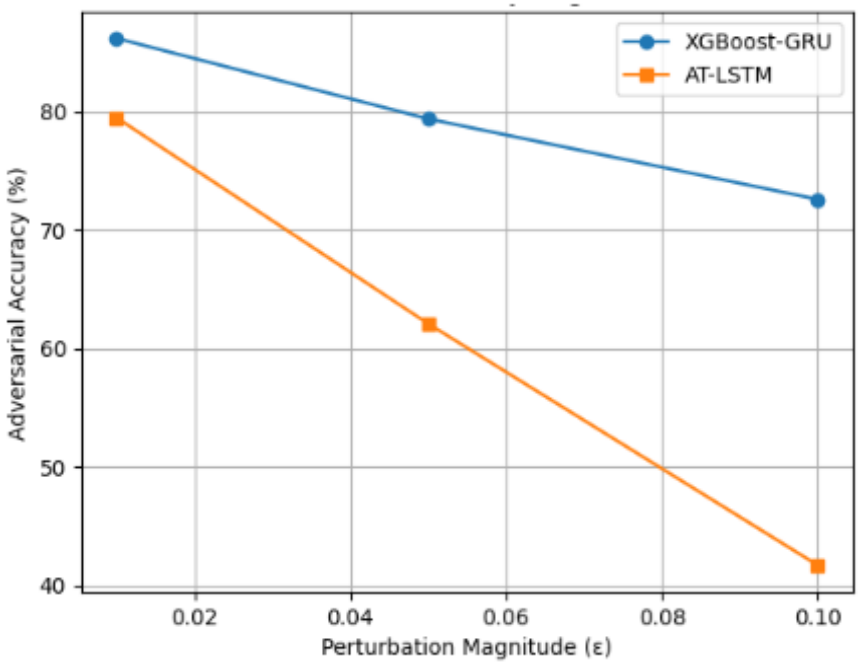

Figure 6: PGD Attack accuracy degradation of the 2 models at ε = 0.10

These results support the hypothesis that architectural simplicity contributes to improved stability under sustained adversarial pressure.

4.4 Evasion Effectiveness and Attack Success Rate Analysis

To directly quantify evasion effectiveness, Attack Success Rate (ASR) is examined at the strongest perturbation level (ε = 0.10). Figure 7 compares ASR values for FGSM and PGD attacks across both architectures.

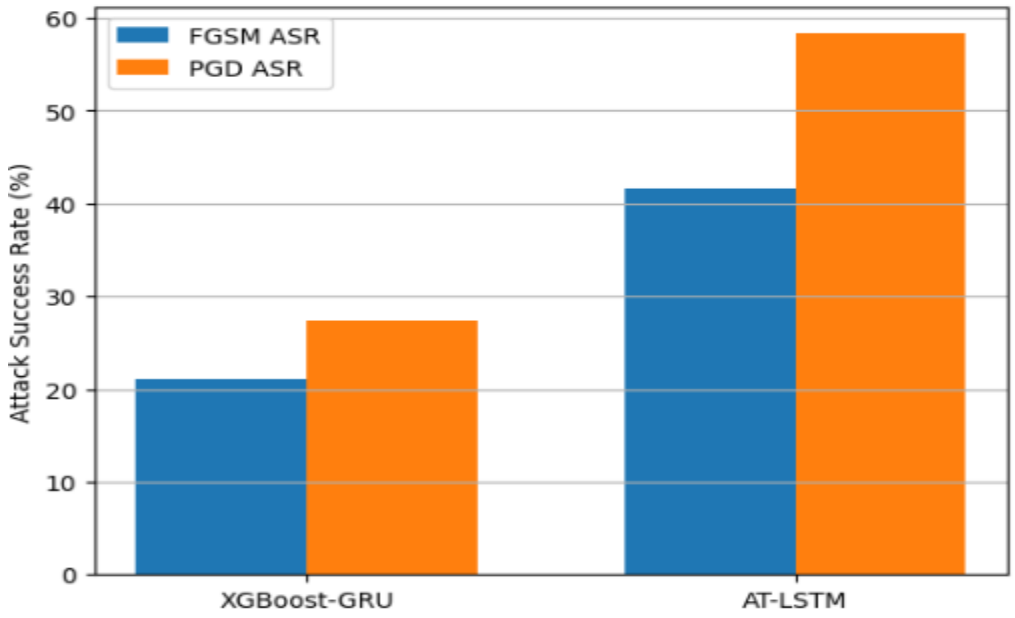

Figure 7: Attack evasion success under strong attack of the 2 Models at ε = 0.10

The AT-LSTM exhibits significantly higher ASR under both attack types of up to 58%, particularly under PGD, where more than half of malicious samples successfully evade detection. On the other hand, the XGBoost–GRU model demonstrates substantially lower ASR (less than 30%), which indicates improved resistance to targeted misclassification. This result confirms that higher adversarial accuracy corresponds to reduced evasion success and reinforces the importance of evaluating IDS models using security-centric metrics rather than accuracy alone.

4.5 Clean vs Adversarial Accuracy Comparison

Figure 8 contrasts clean accuracy with adversarial accuracy under PGD attacks at ε = 0.10. Although the AT-LSTM achieves higher clean accuracy, it experiences noticeable collapse when subjected to adversarial perturbations. The XGBoost–GRU model on the other hand shows smaller discrepancy between clean and adversarial performance.

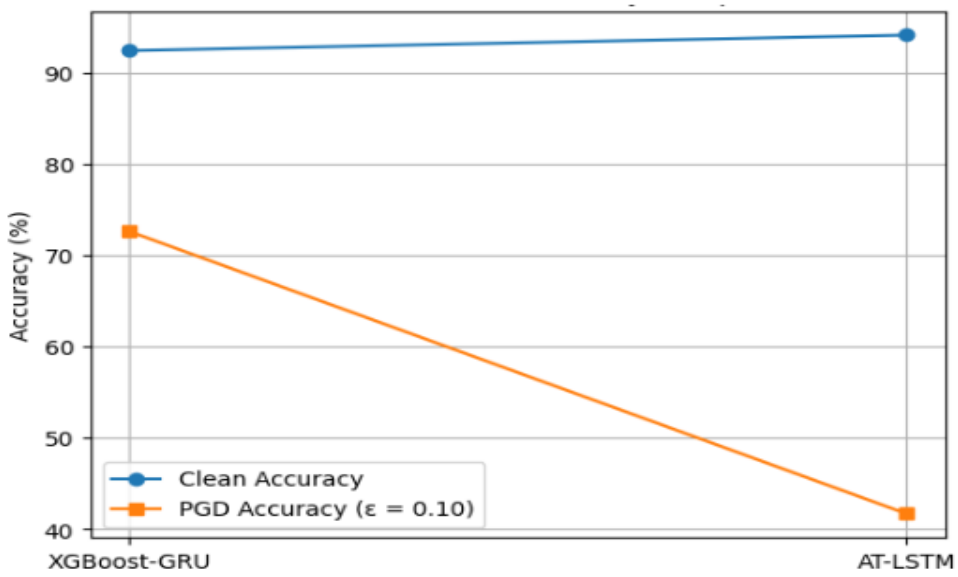

Figure 8: Clean vs Adversarial Comparison of the 2 Models

This comparison highlights a key finding of the study; that clean accuracy is not reliable indicator of adversarial robustness. Models optimized solely for benign performance may exhibit severe vulnerability under hostile conditions.

4.6 Robust Accuracy Retention and Resilience Factor

To normalize robustness across models with differing baseline accuracies, the Resilience Factor (RF) is evaluated under maximum perturbation. Figure 9 presents RF values for both architectures.

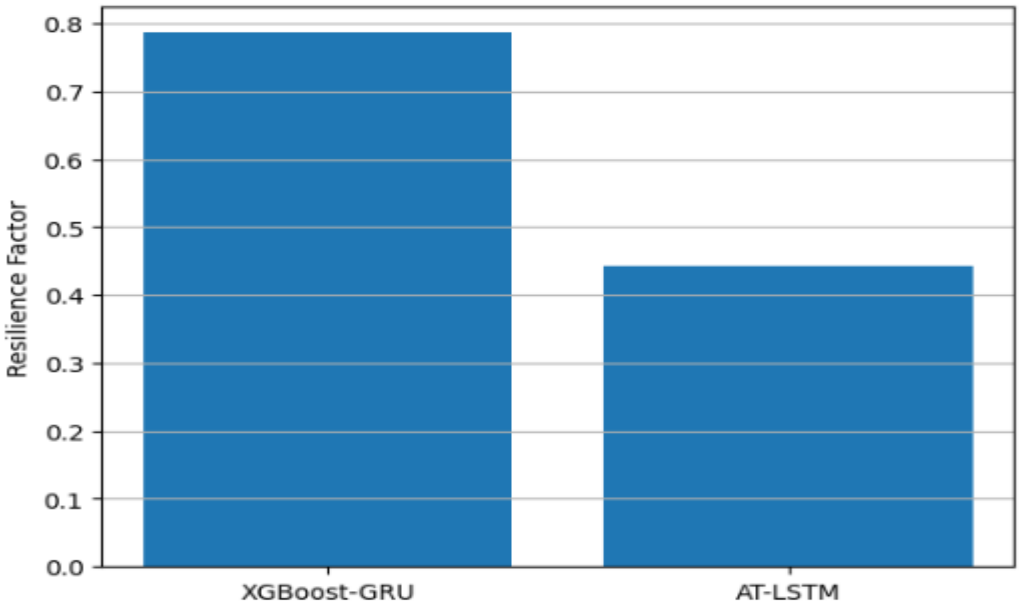

Figure 9: Robust Accuracy Retention of the 2 Models at ε = 0.10

As illustrated in Figure 9, XGBoost–GRU model retains a substantially higher proportion of its baseline accuracy compared to the AT-LSTM. This result demonstrates that lightweight architecture offers superior robust accuracy retention under worst-case evasion. These results support the central claim that reduced complexity can enhance adversarial resilience.

4.7 Analysis of Results

The experimental results presented in Table 1 collectively demonstrate that architectural complexity and adversarial robustness are not positively correlated in intrusion detection systems. Although the attention-based LSTM achieves the highest clean accuracy under benign conditions (Figure 4), this advantage diminishes rapidly when adversarial perturbations are introduced. The observed collapse in detection performance under both FGSM and PGD attacks confirms that clean accuracy alone is unreliable proxy for security in adversarial environments, which is a limitation increasingly acknowledged in adversarial learning literature (Mienye and Swart, 2024)

Table 1: Analysis of Results

Metric	XGBoost–GRU	AT-LSTM
Clean Accuracy (%)	93.0	95.0
FGSM Accuracy (%) (ε = 0.10)	90.0	86.0
PGD Accuracy (%) (ε = 0.10)	72.6	41.7
Accuracy Drop (Clean → PGD) (%)	−20.4	−53.3
FGSM Attack Success Rate (%)	21.1	41.7
PGD Attack Success Rate (%)	27.4	58.3
Resilience Factor (RF)	0.78	0.44

Under FGSM attacks, both models exhibit expected degradation as perturbation magnitude increases. However, the rate of degradation differs substantially between architectures. The XGBoost–GRU model retains approximately 90% accuracy at ε = 0.10 whereas the AT-LSTM degrades more sharply to approximately 86% (Figure 5). This divergence suggests that reduced feature dimensionality and simplified recurrent gating reduce the effectiveness of single-step gradient manipulation. Similar behaviour has been observed in prior studies showing that feature compression can act as an implicit regularizer against adversarial noise (Alharthi et al, 2025). The contrast becomes more visible under PGD attacks which represent stronger and more adaptive adversarial strategy. As shown in Figure 6, the AT-LSTM experiences severe performance collapse with adversarial accuracy dropping to approximately 43% at ε = 0.10. In comparison, the XGBoost–GRU model maintains adversarial accuracy above 70% under the same conditions. The widening robustness gap under iterative attacks indicates that while attention mechanisms is effective for benign sequence modelling, it may expose richer gradient information that adversaries can exploit over multiple optimization steps (Dash, et al., 2024). This result challenges the common assumption that higher-capacity models provide stronger security guarantees.

Attack Success Rate (ASR) analysis further reinforces this conclusion. Figure 7 shows that more than half of malicious samples successfully evade detection in the AT-LSTM under PGD attacks, whereas the XGBoost–GRU model limits successful evasions to less than 30%. This substantial reduction in evasion effectiveness highlights the importance of evaluating IDS performance using security-oriented metrics rather than aggregate accuracy alone. Lower ASR values directly translate into stronger operational reliability in real-world deployment scenarios, especially in resource-constrained IoT environments (Akif et al., 2025). The clean versus adversarial accuracy comparison in Figure 8 provides further additional insight into robustness stability. While the AT-LSTM benefits from marginally higher clean accuracy, its sharp decline under adversarial pressure results in larger accuracy gap between benign and hostile conditions. In contrast, the XGBoost–GRU model exhibits smaller discrepancy, which indicates more consistent behaviour across threat scenarios. This stability is quantitatively captured by the Resilience Factor metric shown in Figure 9. The higher RF value of the lightweight model confirms superior accuracy retention under worst-case perturbation and enables normalized comparison across architectures with different baseline performance.

5. CONCLUSION

This study investigated the adversarial resilience of intrusion detection systems in IoT environments by focusing on detection stability under evasion attacks rather than peak accuracy under benign conditions. The developed XGBoost–GRU model was compared against high-complexity attention-based LSTM using gradient-based evasion attacks through controlled adversarial evaluation framework. Experimental results consistently demonstrate that the lightweight hybrid architecture maintains higher adversarial accuracy, lower attack success rates and superior robust accuracy retention under both FGSM and PGD attacks. The findings revealed that even though complex attention-based architectures may achieve marginally higher clean accuracy, they exhibit noticeable vulnerability under adversarial manipulation. On the other hand, feature reduction and simplified temporal modelling contribute to improved robustness, particularly under extreme attacks.

5.1 Limitations and Future Work

While this study provides strong evidence of the robustness benefits of lightweight architectures, several limitations remain. First, the evaluation focuses exclusively on white-box evasion attacks; transferability under black-box or adaptive threat models was not examined. Second, perturbations were restricted to numerical traffic features, and protocol-level or semantic feature manipulation was

outside the scope of this work. Third, experiments were conducted on single benchmark dataset which may limit generalizability across heterogeneous network environments.

Future research should address these limitations by extending evaluation to black-box and transfer-based attacks, incorporating protocol-aware adversarial constraints, and validating robustness across multiple real-world IoT datasets. Additionally, integrating lightweight adversarial training and adaptive feature selection mechanisms represents a promising direction for further improving IDS resilience without increasing model complexity.

REFERENCES

Aleesa, A., Younis, M. O. H. A. M. M. E. D., Mohammed, A. A., & Sahar, N. (2021). Deep-intrusion detection system with enhanced UNSW-NB15 dataset based on deep learning techniques. *Journal of Engineering Science and Technology*, *16*(1), 711-727.

Ali, M., Shahroz, M., Mushtaq, M. F., Alfarhood, S., Safran, M., & Ashraf, I. (2024). Hybrid machine learning model for efficient botnet attack detection in IoT environment. *IEEe Access*, *12*, 40682-40699.

Alnuaimi, A. F., & Albaldawi, T. H. (2024). An overview of machine learning classification techniques. In *BIO Web of Conferences* (Vol. 97, p. 00133). EDP Sciences.

Alsharaiah, M. A., Abu-Shareha, A. A., Abualhaj, M., Baniata, L. H., Al-saaidah, A., Kharma, Q. M., & Al-Zyoud, M. M. (2024). An innovative network intrusion detection system (NIDS): Hierarchical deep learning model based on Unsw-Nb15 dataset. *International Journal of Data & Network Science*, *8*(2).

Asharf, J., Moustafa, N., Khurshid, H., Debie, E., Haider, W., & Wahab, A. (2020). A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions. *Electronics*, *9*(7), 1177.

Armijos, A., & Cuenca, E. (2023, November). Zero-day attacks: review of the methods used based on intrusion detection and prevention systems. In *2023 IEEE Colombian Caribbean Conference (C3)* (pp. 1-6). IEEE.

Alharthi, M., Medjek, F., & Djenouri, D. (2025). Ensemble learning approaches for multi-class intrusion detection systems for the internet of vehicles (IoV): a comprehensive survey. *Future Internet*, *17*(7), 317.

Alanazi, R., & Aljuhani, A. (2023). Anomaly Detection for Industrial Internet of Things Cyberattacks. *Computer Systems Science & Engineering*, *44*(3).

Asharf, J., Moustafa, N., Khurshid, H., Debie, E., Haider, W., & Wahab, A. (2020). A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions. *Electronics*, *9*(7), 1177.

Alwahedi, F., Aldhaheri, A., Ferrag, M. A., Battah, A., & Tihanyi, N. (2024). Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models. *Internet of Things and Cyber-Physical Systems*, *4*, 167-185.

Akif, M. A., Butun, I., Williams, A., & Mahgoub, I. (2025). Hybrid machine learning models for intrusion detection in iot: Leveraging a real-world iot dataset. *arXiv preprint arXiv:2502.12382*.

Alraba'nah, Y., Al-Sharaeh, S., & Al Hindi, G. (2025). Enhancing intrusion detection using hybrid long short-term memory and XGBoost. *Journal of Soft Computing and Data Mining*, *6*(1), 247-261.

Ahmed, S., Raza, B., Hussain, L., Aldweesh, A., Omar, A., Khan, M. S., ... & Nadim, M. A. (2023). The deep learning resnet101 and ensemble xgboost algorithm with hyperparameters optimization accurately predict the lung cancer. *Applied Artificial Intelligence*, *37*(1), 2166222.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.

Dash, S. K., Dash, S., Mahapatra, S., Mohanty, S. N., Khan, M. I., Medani, M., ... & Gupta, M. (2024). Enhancing DDoS attack detection in IoT using PCA. *Egyptian Informatics Journal*, *25*, 100450.

Halbouni, A., Gunawan, T. S., Habaebi, M. H., Halbouni, M., Kartiwi, M., & Ahmad, R. (2022). Machine learning and deep learning approaches for cybersecurity: A review. *IEEE Access*, *10*, 19572-19585.

Khaw, Y. M., Jahromi, A. A., Arani, M. F., Sanner, S., Kundur, D., & Kassouf, M. (2020). A deep learning-based cyberattack detection system for transmission protective relays. *IEEE Transactions on Smart Grid*, *12*(3), 2554-2565.

Kasongo, S. M., & Sun, Y. (2020). Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. *Journal of Big Data*, *7*(1), 105.

Kasongo, S. M. (2023). A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework. *Computer Communications*, *199*, 113-125.

Le, T. T. H., Oktian, Y. E., & Kim, H. (2022). XGBoost for imbalanced multiclass classification-based industrial internet of things intrusion detection systems. *Sustainability*, *14*(14), 8707.

Layeghy, S., Baktashmotlagh, M., & Portmann, M. (2023). DI-NIDS: Domain invariant network intrusion detection system. *Knowledge-Based Systems*, *273*, 110626.

Liu, G., Zhang, W., Wang, X., King, S., & Yu, S. (2024). A membership inference and adversarial attack defense framework for network traffic classifiers. *IEEE Transactions on Artificial Intelligence*, *6*(2), 317-332.

Mienye, I. D., and Swart, T. G. (2024). A comprehensive review of deep learning: Architectures, recent advances, and applications. *Information*, *15*(12), 755.

Meena, G., and Indian, A. (2025, October). IDS-IoT: Intrusion Detection System for the Internet of Things Using Enhanced Long-Short Term Memory. In *Artificial Intelligence and Applications*.

Moustafa, N., and Slay, J. (2015, November). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 military communications and information systems conference (MilCIS)* (pp. 1-6). Ieee.

Meena, G., and Choudhary, R. R. (2017, July). A review paper on IDS classification using KDD 99 and NSL KDD dataset in WEKA. In *2017 International Conference on Computer, Communications and Electronics (Comptelix)* (pp. 553-558). IEEE.

More, S., Idrissi, M., Mahmoud, H., and Asyhari, A. T. (2024). Enhanced intrusion detection systems performance with UNSW-NB15 data analysis. *Algorithms*, *17*(2), 64.

Nosouhian, S., Nosouhian, F., and Khoshouei, A. K. (2021). A review of recurrent neural network architecture for sequence learning: Comparison between LSTM and GRU

Rahman, M. M., Al Shakil, S., and Mustakim, M. R. (2025). A survey on intrusion detection system in IoT networks. *Cyber Security and Applications*, *3*, 100082.

Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, *2*(6), 1-20.

Soon, H. F., Amir, A., Nishizaki, H., Zahri, N. A. H., Kamarudin, L. M., & Azemi, S. N. (2024). Evaluating Tree-based Ensemble Strategies for Imbalanced Network Attack Classification. *International Journal of Advanced Computer Science & Applications*, *15*(1).

Sarhan, M., Layeghy, S., Moustafa, N., and Portmann, M. (2020, December). Netflow datasets for machine learning-based network intrusion detection systems. In *International Conference on*.

Zhang, Y., Gandhi, Y., Li, Z., & Xiao, Z. (2022, September). Improving the classification effectiveness of network intrusion detection using ensemble machine learning techniques and deep neural networks. In *2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)* (pp. 117-123). IEEE.

Vitorino, J., Silva, M., Maia, E., & Praça, I. (2025). Reliable feature selection for adversarially robust cyber-attack detection. *Annals of Telecommunications*, *80*(3), 341-355. doi: https://doi.org/10.1007/s12243-024-01047-z..