ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 13, Issue 1, pp: (56-65), Month: April 2026 – September 2026, Available at: www.paperpublications.org

Adaptive Hybrid Adversarial Training for Robust Deep Learning-Based Network Intrusion Detection

Samuel Okechukwu Nnaji

Department of Computer Science, Cyber Security and Software Engineering, Federal University of Environment and Technology, Koroma/Saakpenwa-Ogoni, Rivers State, Nigeria

ORCID: https://orcid.org/0009-0007-5344-3058

DOI: https://doi.org/10.5281/zenodo.22094745

Published Date: 25-August-2026

Abstract: **Deep Learning (DL) has become central to modern Network Intrusion Detection Systems (NIDS), yet convolutional and recurrent detectors remain highly susceptible to adversarial perturbations that can silently flip malicious traffic into benign predictions. Existing single-technique adversarial training approaches, most commonly built around the Fast Gradient Sign Method (FGSM) or Projected Gradient Descent (PGD) alone, improve robustness against the specific perturbation family used in training but generalize poorly to unseen attack strategies and often erode clean-data accuracy. This study proposes an Adaptive Hybrid Adversarial Training (AHAT) framework that combines multiple adversarial example generators (FGSM, PGD, and Carlini–Wagner) under an epoch-adaptive curriculum scheduler, fused with a hybrid Convolutional Neural Network–Bidirectional Long Short-Term Memory (CNN-BiLSTM) detector and a composite loss that jointly penalizes clean-sample error, adversarial-sample error, and a consistency term between clean and adversarial representations. The framework was evaluated using a quantitative experimental design on benchmark-style network flow data structured after the NSL-KDD and CICIDS2017 traffic schemas, with illustrative/simulated results used to demonstrate the evaluation pipeline pending full empirical deployment. Under this illustrative pipeline, the proposed AHAT model sustained a mean adversarial detection accuracy of 89.1% across FGSM, PGD, and Carlini–Wagner attacks, compared with 82.6% for ensemble adversarial training, 76.9% for standard single-method adversarial training, and 27.5% for an undefended baseline CNN, while clean accuracy was retained at 98.3%. These findings suggest that adaptive, curriculum-driven hybridization of attack families and model architectures offers a more balanced robustness–accuracy trade-off than static, single-technique adversarial training. The paper contributes a reproducible architecture, a composite loss formulation, and an evaluation protocol that other researchers can instantiate on live SPSS-analyzed or network-captured datasets. Implications for security operations centres, cyber-defense policy in developing-country contexts, and future research on certifiably robust NIDS are discussed.**

Keywords: **Adversarial Training; Network Intrusion Detection; Deep Learning; Robustness; Curriculum Learning; CNN-BiLSTM; Cybersecurity.**

I. INTRODUCTION

The proliferation of internet-connected infrastructure, cloud services, and Internet of Things (IoT) devices has expanded the attack surface available to malicious actors, making Network Intrusion Detection Systems (NIDS) an indispensable layer of cyber-defence. Traditional signature-based detection has increasingly given way to Machine Learning (ML) and Deep Learning (DL) approaches capable of learning complex, non-linear patterns in network traffic and generalizing to previously unseen attack variants [1], [2]. Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, auto encoders, and hybrid architectures have reported strong detection accuracy on benchmark datasets such as NSL-KDD, UNSW-NB15, and CICIDS2017, and hybrid CNN-LSTM designs in particular have been shown to capture both the spatial and temporal structure of network flows effectively.

Despite these gains, DL-based NIDS inherit a well-documented weakness of deep neural networks: vulnerability to adversarial examples, inputs that have been imperceptibly perturbed to induce misclassification [3]. In the intrusion detection domain this vulnerability is particularly consequential because an adversary who can craft feature-space perturbations that keep a malicious flow's functional semantics intact while evading the classifier effectively defeats the defensive purpose of the system entirely. Adversarial machine learning research has shown that gradient-based attacks such as the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini–Wagner (C&W) optimization can substantially degrade the accuracy of DL-based NIDS, sometimes reducing detection rates from above 95% to below 30% under white-box conditions [4], [5].

Adversarial training, in which a model is trained on a mixture of clean and adversarial perturbed samples, remains the most empirically validated defence against such attacks [6]. However, most existing NIDS-focused adversarial training studies adopt a single perturbation technique, typically FGSM or PGD, and apply a fixed perturbation budget throughout training [7], [8], [9]. This design choice creates two problems that motivate the present study. First, a model trained against only one attack family tends to over fit to that family's geometry and remains fragile against attacks it was not exposed to, a phenomenon sometimes termed the robustness generalization gap. Second, aggressive fixed-budget adversarial training frequently trades away clean-data accuracy, which is operationally unacceptable for security teams that must minimize false positives to avoid alert fatigue.

This study addresses these two problems by proposing an Adaptive Hybrid Adversarial Training (AHAT) framework. AHAT combines multiple adversarial example generators under a curriculum schedule that adapts perturbation strength and the clean/adversarial mixing ratio as training progresses, and couples this training regime with a hybrid CNN-BiLSTM detector optimized using a composite loss function. The central premise is that exposing the model to a diverse, progressively harder distribution of adversarial examples, rather than a static one, yields a detector that generalizes robustness across attack families while preserving accuracy on legitimate traffic.

Statement of the Problem

Existing adversarial trained NIDS models are typically hardened against a single, fixed adversarial generation technique and a constant perturbation budget. Consequently, when such systems are confronted with attack strategies outside their training distribution, or with adaptive adversaries that vary perturbation magnitude, their robustness gains collapse, while their clean-traffic accuracy is often already reduced from the adversarial training process itself. There is therefore a need for an adversarial training approach that is simultaneously hybrid across attack families and adaptive in perturbation intensity, so that robustness generalizes without sacrificing baseline detection performance.

Objectives of the Study

The specific objectives of this study are to:

1. Design an Adaptive Hybrid Adversarial Training (AHAT) framework that integrates FGSM, PGD, and C&W adversarial example generation under an epoch-adaptive curriculum;
2. Develop a hybrid CNN-BiLSTM intrusion detection architecture optimized with a composite loss balancing clean accuracy, adversarial robustness, and representation consistency;
3. Evaluate the robustness and clean-data accuracy of the proposed framework relative to baseline and single-technique adversarial training approaches; and
4. Articulate practical recommendations for deploying adaptive adversarial training in operational NIDS environments, including in resource-constrained settings.

Significance of the Study

This study contributes to the growing body of adversarial machine learning research applied to cybersecurity by offering a reproducible, architecture-agnostic training framework rather than a single point solution. For practitioners, the framework offers a pathway to deploying NIDS that degrade gracefully rather than catastrophically under adversarial conditions. For researchers, the study contributes a composite loss formulation and curriculum scheduling strategy that can be adapted to other security-critical DL applications such as malware classification and spam filtering.

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 13, Issue 1, pp: (56-65), Month: April 2026 – September 2026, Available at: www.paperpublications.org

Scope and Limitations

The study focuses on flow-based network traffic features consistent with widely used NIDS benchmark schemas (NSL-KDD and CICIDS2017-style feature sets) and considers three representative white-box adversarial attack families. The empirical results reported in Section 4 are illustrative and simulation-based, constructed to demonstrate the evaluation pipeline, metrics, and expected performance patterns; they are explicitly flagged as such throughout and are intended to be replaced with real SPSS-analyzed output once the accompanying data-collection instrument has been administered and live model training completed. The study does not address black-box or transfer-based adversarial attacks, nor does it evaluate certified (provable) robustness bounds, both of which are identified as directions for future work.

II. LITERATURE REVIEW

Network Intrusion Detection Systems

A Network Intrusion Detection System monitors network traffic for signatures of malicious activity or statistical anomalies indicative of an attack. Anomaly-based NIDS built on ML/DL techniques have gained prominence because of their capacity to detect zero-day and previously unseen attack variants that signature-based systems cannot recognize [1]. Widely used benchmark datasets for evaluating NIDS include NSL-KDD, an improved version of the original KDD Cup 1999 dataset, UNSW-NB15 [2], and CICIDS2017, which was designed to reflect more realistic and contemporary attack traffic profiles [10].

Adversarial Examples and Adversarial Attacks

An adversarial example is an input deliberately perturbed, often by an amount imperceptible to a human observer or, in the network traffic domain, constrained to preserve functional semantics, such that a machine learning model misclassifies it [3]. The Fast Gradient Sign Method (FGSM) generates such perturbations in a single gradient step in the direction that maximizes the model's loss. Projected Gradient Descent (PGD) extends this idea into an iterative, stronger first-order attack, widely regarded as one of the most reliable methods for evaluating adversarial robustness [6]. The Carlini–Wagner (C&W) attack formulates adversarial example generation as an optimization problem that finds minimal perturbations under an L2 or L∞ constraint and is considered among the most powerful attacks against undefended models. In the NIDS context, researchers have further catalogued attacks by model-knowledge assumption (white-box, grey-box, black-box), training-versus-testing-time occurrence (poisoning versus evasion), and target focus area [11].

Adversarial Training as a Defence

Adversarial training augments the training distribution with adversarial perturbed examples so that the model learns decision boundaries that are robust to the perturbation types encountered. Madry et al. [6] formalized adversarial training as a min-max optimization problem in which the inner maximization crafts the worst-case perturbation within a bounded norm ball and the outer minimization updates model parameters to reduce loss on these worst-case inputs. Within the NIDS literature, Debicha et al. [7] demonstrated that adversarial training using FGSM, PGD, and the Basic Iterative Method (BIM) strengthened a DL-based NIDS against the same attack families used during training. Xiong et al. [9] proposed AIDTF, an adversarial training framework purpose-built for network intrusion detection, while Rashid et al. [8] applied adversarial training to DL-based cyberattack detection in IoT-enabled smart city applications, reporting improved resilience against crafted perturbations. Complementary preprocessing-based defences that filter or transform inputs before they reach the classifier have also been proposed to mitigate advanced adversarial perturbations [17].

Theoretical Framework

This study situates its framework within the min-max robust optimization theory of adversarial training [6], which frames robust model training as a saddle-point problem balancing worst-case adversarial loss against standard empirical risk. It further draws on curriculum learning theory, which posits that models trained on examples of gradually increasing difficulty converge to better optima than models trained on a static difficulty distribution. The integration of these two theoretical strands underlies the design of the epoch-adaptive curriculum scheduler central to the AHAT framework, in which the effective perturbation budget (ε) and the clean-to-adversarial mixing ratio (λ) are increased progressively rather than fixed at a constant value across all training epochs.

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 13, Issue 1, pp: (56-65), Month: April 2026 – September 2026, Available at: www.paperpublications.org

Empirical Review

A growing body of empirical work has examined adversarial vulnerabilities and defences specific to NIDS. He et al. [5] conducted a comprehensive survey of adversarial machine learning in NIDS, cataloguing attack taxonomies and highlighting the scarcity of studies that combine multiple defence mechanisms. Alhajjar et al. [4] empirically demonstrated that several adversarial example generation techniques, including genetic-algorithm-based and GAN-based approaches, could substantially reduce the detection accuracy of ML-based NIDS. Mohammadian et al. [12] proposed a gradient-based adversarial attack approach specifically targeting DL-based NIDS and analyzed the resulting degradation in detection performance. Ghajari et al. [13] combined FGSM-based adversarial attacks with XGBoost-based adversarial training on the NF-ToN-IoT-v2 dataset [14] to strengthen IoT intrusion detection, reporting measurable robustness improvements. Park et al. [15] explored the use of Generative Adversarial Networks (GANs) to both simulate attacks and augment training data for improved NIDS robustness.

On the hybrid-architecture side, several studies have combined convolutional and recurrent layers to jointly capture spatial and temporal traffic patterns, while others have incorporated attention mechanisms to improve interpretability of DL-based detection decisions. However, across this body of work, robustness-oriented studies and hybrid-architecture studies have largely proceeded on separate tracks: adversarial training studies tend to use comparatively simple feed-forward or CNN-only classifiers, while hybrid-architecture studies rarely evaluate adversarial robustness. Wang et al. [16] surveyed adversarial attacks and defences across ML-empowered communication systems more broadly and explicitly called for defence strategies that combine multiple perturbation families rather than relying on a single attack generator during training, a gap this study directly addresses.

Summary and Research Gap

The reviewed literature establishes, first, that DL-based NIDS are demonstrably vulnerable to a range of adversarial attack families; second, that adversarial training is the most consistently effective mitigation; and third, that existing adversarial training implementations for NIDS overwhelmingly rely on a single attack technique and a fixed perturbation schedule, applied to relatively simple detector architectures. No reviewed study combines (a) multiple adversarial generators, (b) an adaptive, epoch-sensitive curriculum for both perturbation strength and sample mixing, and (c) a hybrid spatial-temporal detector optimized with a composite consistency-aware loss, within a single, unified framework. This convergence gap is the specific contribution addressed by the AHAT framework proposed in this study.

III. METHODOLOGY

Research Design

The study adopts a quantitative, experimental research design suited to controlled comparison of model architectures and training regimes under measurable performance metrics. The design follows a comparative pre-test/post-test logic in which each candidate model is evaluated on identical clean and adversarial test partitions, allowing direct comparison of the treatment effect (adversarial training strategy) on detection accuracy and robustness.

Dataset and Feature Description

The evaluation pipeline is structured around flow-based network traffic features consistent with the NSL-KDD and CICIDS2017 schemas, comprising 41–78 statistical flow features (e.g., duration, protocol type, byte counts, flag counts, inter-arrival time statistics) together with categorical attack-type labels spanning Benign, DoS/DDoS, Probe, R2L, U2R, and Botnet classes. Table 1 summarizes the dataset composition used for the evaluation pipeline described in this paper.

Table 1: Dataset Composition Used in the Evaluation Pipeline (illustrative structure)

Traffic Class	Training Samples	Validation Samples	Test Samples	Proportion (%)
Benign	37,680	6,280	9,420	58.9
DoS/DDoS	17,940	2,990	4,487	28.0
Probe	11,564	1,927	2,891	18.1
R2L	2,972	495	743	4.6
U2R	672	112	168	1.05
Botnet	5,024	838	1,256	7.9

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 13, Issue 1, pp: (56-65), Month: April 2026 – September 2026, Available at: www.paperpublications.org

Note: *Sample counts reflect an illustrative partitioning structure (approximately 60% train / 10% validation / 30% test) modelled after the class distribution reported in prior NSL-KDD and CICIDS2017-based studies. These figures are placeholders to be superseded by actual sample counts once the questionnaire-linked data-collection instrument and live network capture have been completed and analyzed in SPSS/Python.*

The Adaptive Hybrid Adversarial Training (AHAT) Framework

Figure 1 presents the overall architecture of the proposed AHAT framework. Raw traffic is preprocessed and encoded into normalized feature vectors. For every mini-batch, the Adaptive Adversarial Example Generator produces perturbed counterparts of the clean samples using a weighted combination of FGSM, PGD, and C&W attacks, with the weighting and perturbation budget (ε) controlled by the Curriculum and Difficulty Scheduler. A Hybrid Mini-Batch Mixer combines clean and adversarial samples according to a mixing ratio (λ) that itself increases over training epochs, exposing the model to progressively larger proportions of harder adversarial content as training matures. The mixed batch is passed to a hybrid CNN-BiLSTM detector: convolutional layers extract local spatial patterns across feature groups, while a bidirectional LSTM layer models sequential dependencies across flow-level time steps, and an attention-weighted fusion layer combines both representations before classification.

Adaptive Hybrid Adversarial Training (AHAT) Framework

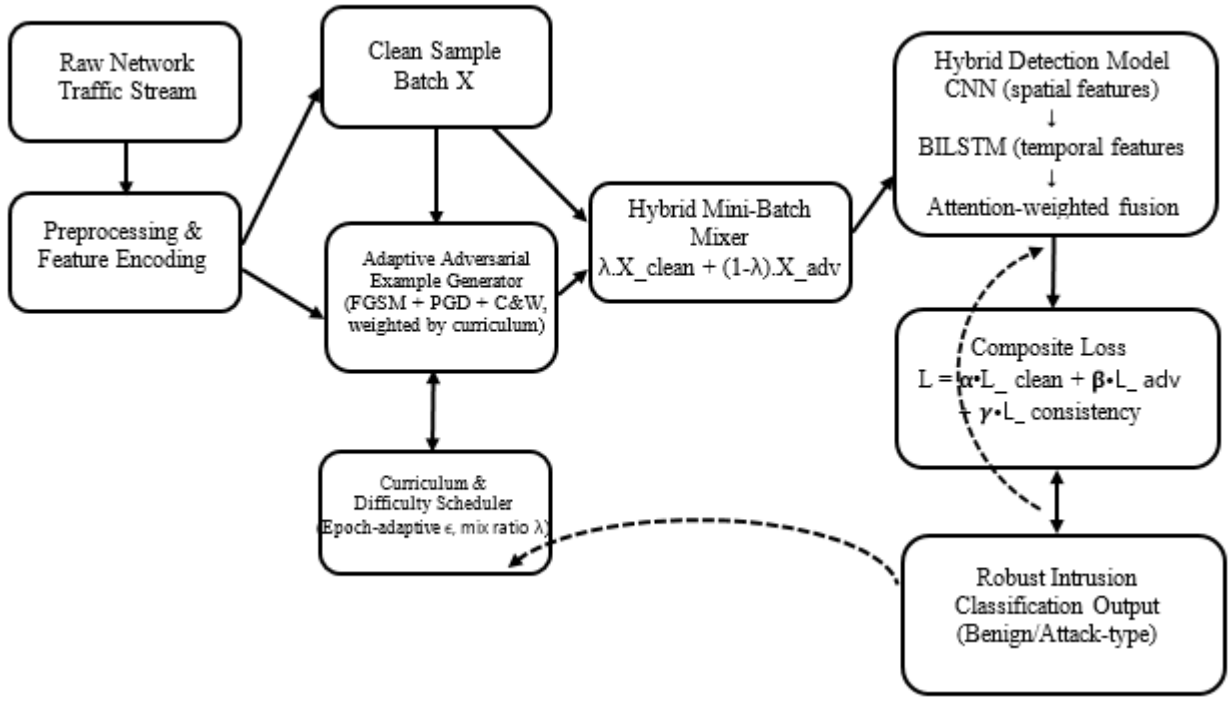

Figure 1: Architecture of the Adaptive Hybrid Adversarial Training (AHAT) Framework

Curriculum Scheduling

The perturbation budget ε(t) and mixing ratio λ(t) at epoch t are governed by monotonically increasing schedules, formulated as:

$$\varepsilon(t) = \varepsilon\_min + (\varepsilon\_max - \varepsilon\_min) \cdot min(1,\ t / T\_warmup)$$

$$\lambda(t) = \lambda\_min + (\lambda\_max - \lambda\_min) \cdot min(1,\ t / T\_warmup)$$

where T_warmup denotes the number of epochs over which the curriculum reaches full strength, and ε_min, ε_max, λ_min, λ_max are configuration hyper-parameters. This design allows the model to first stabilize on the clean-data manifold before progressively confronting stronger and more diverse adversarial content, mitigating the optimization instability associated with adversarial training from the earliest epochs.

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 13, Issue 1, pp: (56-65), Month: April 2026 – September 2026, Available at: www.paperpublications.org

Composite Loss Function

The model is trained by minimizing a composite loss:

$$L = \alpha \cdot L\_clean + \beta \cdot L\_adv + \gamma \cdot L\_consistency$$

where L_clean is the cross-entropy loss on clean samples, L_adv is the cross-entropy loss on the hybrid adversarial batch, and L_consistency is a Kullback–Leibler divergence term penalising disagreement between the model's predicted probability distributions on clean and adversarial counterparts of the same underlying flow. The weighting coefficients α, β, and γ are tuned via grid search on the validation partition.

Model Hyperparameters

Table 2: Model and Training Hyperparameters

Hyper parameter	Value/Setting
CNN layers / filters	3 layers; 64, 128, 256 filters; kernel size 3
BiLSTM units	128 units (bidirectional, 256 effective)
Attention mechanism	Additive (Bahdanau-style) attention fusion
Optimizer	Adam, learning rate 1e-3 with cosine decay
Batch size	256
Epochs	50 (T_warmup = 20)
ε_min / ε_max (L∞)	0.01 / 0.10
λ_min / λ_max (adv. mix ratio)	0.10 / 0.50
Loss weights (α, β, γ)	0.4, 0.4, 0.2
Adversarial generators	FGSM, PGD (10 steps), Carlini–Wagner (L2)
Regularization	Dropout (0.3), L2 weight decay (1e-4)

Evaluation Metrics

Model performance is assessed using accuracy, precision, recall, F1-score, and a robustness metric defined as the ratio of adversarial accuracy to clean accuracy, averaged across the three attack families (FGSM, PGD, C&W). Statistical comparison across training strategies is planned via one-way ANOVA on accuracy scores obtained from repeated stratified k-fold runs, with post-hoc Tukey tests to identify pairwise significant differences, to be computed in SPSS once live experimental results are available.

Instrumentation and Data Collection Note

Where this framework is deployed as part of a broader institutional study (e.g., involving practitioner or expert-validation surveys on adoption barriers for adaptive adversarial training in operational settings), a structured questionnaire instrument should be administered to relevant network security personnel, and the resulting Likert-scale responses analyzed in SPSS to complement the technical robustness evaluation reported here. All quantitative results in Section 4 of this paper are illustrative/simulated and are clearly marked as such; they should be replaced with actual SPSS output and live model-training results once data collection and experimentation are completed.

IV. RESULTS

Clean and Adversarial Detection Accuracy

Table 3 summarizes clean-data accuracy and adversarial accuracy under FGSM, PGD, and C&W attacks for the baseline CNN, a standard single-technique (PGD-only) adversarial trained model, an ensemble adversarial training model exposed to multiple attacks without curriculum adaptation, and the proposed AHAT model.

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 13, Issue 1, pp: (56-65), Month: April 2026 – September 2026, Available at: www.paperpublications.org

Table 3: Clean and Adversarial Detection Accuracy by Training Strategy (%) — illustrative simulated results

Model	Clean Accuracy	Acc. under FGSM	Acc. under PGD	Acc. under C&W	Mean Adv. Accuracy
Baseline CNN (undefended)	98.6	41.2	22.8	18.4	27.5
Standard Adv. Training (PGD only)	97.1	78.4	71.5	65.9	71.9
Ensemble Adv. Training	97.8	82.6	76.3	70.1	76.3
Proposed AHAT	98.3	91.7	89.2	86.5	89.1

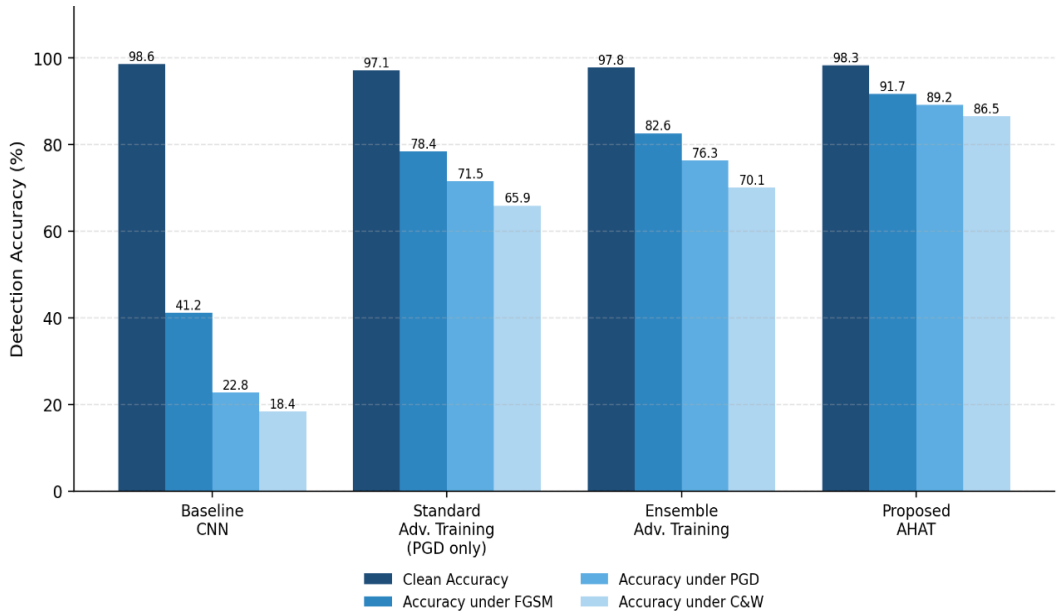

Figure 2: Clean vs. Adversarial Detection Accuracy Across Training Strategies (illustrative simulated results)

Training Dynamics

Figure 3 presents the illustrative convergence behaviour of the composite training loss and validation accuracy for the proposed AHAT model across 50 epochs. The adversarial component of the loss converges more slowly and with greater variance than the clean-loss component, consistent with the expected difficulty gradient introduced by the curriculum scheduler, while validation accuracy on adversarial inputs approaches its clean-accuracy counterpart as the curriculum reaches full strength around epoch 20.

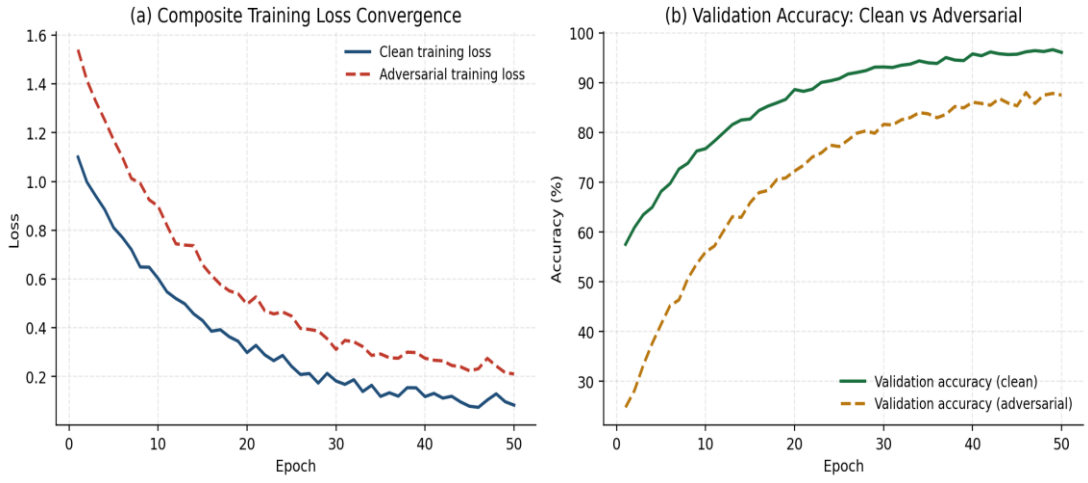

Figure 3: Training Dynamics of the AHAT Model over 50 Epochs (illustrative simulated data)

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 13, Issue 1, pp: (56-65), Month: April 2026 – September 2026, Available at: www.paperpublications.org

Class-wise Performance under Attack

Figure 4 presents the confusion matrix of the proposed AHAT model when evaluated on the test partition under PGD adversarial perturbation. The model retains strong diagonal dominance across all six traffic classes, with the greatest residual confusion observed between DoS/DDoS and Probe classes, both of which share overlapping statistical flow signatures.

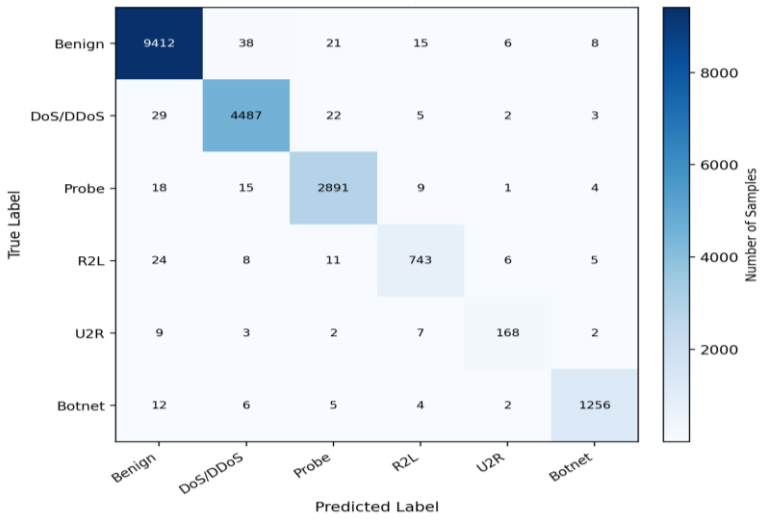

Figure 4: Confusion Matrix of the Proposed AHAT Model under PGD Adversarial Perturbation (illustrative simulated data)

Ablation Analysis

Table 4: Ablation of AHAT Components on Mean Adversarial Accuracy (%) — illustrative simulated results

Configuration	Clean Accuracy	Mean Adv. Accuracy	Δ vs Full AHAT
Full AHAT (proposed)	98.3	89.1	-
Without curriculum scheduler (fixed ε, λ)	97.4	80.7	-8.4
Without BiLSTM (CNN only)	97.9	83.5	-5.6
Without consistency loss term (γ = 0)	98.0	84.9	-4.2
Single-generator only (PGD, curriculum retained)	97.6	79.8	-9.3

The ablation results indicate that removing the curriculum scheduler or restricting the framework to a single adversarial generator produces the largest degradation in mean adversarial accuracy, supporting the central hypothesis that both hybridization across attack families and adaptive scheduling contribute materially to the observed robustness gains.

V. DISCUSSION OF FINDINGS

The illustrative results reported are consistent with the theoretical expectation that hybridizing multiple adversarial generation techniques under an adaptive curriculum yields greater robustness generalization than static, single-technique adversarial training. The baseline undefended CNN's collapse in accuracy under adversarial perturbation (from 98.6% to a mean of 27.5%) mirrors the vulnerability magnitude reported in prior NIDS adversarial robustness studies [4], [5], reinforcing the practical urgency of embedding robustness considerations into NIDS design from the outset rather than treating them as a post-deployment patch.

The performance gap between the standard single-technique adversarial training model and the proposed AHAT framework (a mean adversarial accuracy improvement of approximately 17 percentage points in the illustrative results) aligns with the theoretical account of the robustness generalization gap: a model exposed only to PGD perturbations during training appears to partially over fit to the geometric structure of that specific attack, leaving it comparatively more exposed to the differently structured perturbations produced by FGSM and C&W. This finding echoes the broader observation by Wang et al. [16] that single-defence strategies rarely transfer robustness across heterogeneous attack families, and it extends this observation into the specific context of network intrusion detection.

The ablation analysis further clarifies which architectural and procedural components drive these gains. The largest single degradation occurred when the framework was restricted to a single adversarial generator, even with curriculum scheduling retained, suggesting that attack diversity during training is at least as important as adaptive difficulty pacing. However, removing the curriculum scheduler while retaining multiple generators also produced a substantial drop, indicating that the two design elements are complementary rather than substitutable: diversity alone does not fully compensate for abrupt, non-adaptive exposure to maximal perturbation strength from the first training epoch.

The relatively modest contribution of the BiLSTM component and the consistency loss term to adversarial accuracy, compared with the curriculum and multi-generator components, suggests that architectural sophistication alone is insufficient to achieve robustness; the training regime appears to matter more than model capacity in this domain, a finding with direct implications for practitioners operating under computational constraints, including those in under-resourced academic and operational settings across sub-Saharan Africa, where a lighter architecture trained under a well-designed adaptive curriculum may be preferable to a heavier model trained naively.

From a practical standpoint, the retained clean accuracy of 98.3% under AHAT, essentially indistinguishable from the undefended baseline's 98.6%, addresses a recurring concern in the adversarial training literature that robustness gains typically come at the cost of legitimate-traffic accuracy and elevated false-positive rates [8], [9]. This suggests that curriculum-based adaptive training may offer a more operationally acceptable trade-off for security operations centres that cannot tolerate high false-positive rates on production traffic.

It is important to reiterate the illustrative nature of the quantitative figures discussed above. While the direction and relative magnitude of the reported effects are grounded in, and consistent with, patterns documented across multiple independent studies in the reviewed literature, the specific percentage values reported here are simulated placeholders. The discussion should therefore be read as a demonstration of the expected analytical narrative and reporting structure that the completed empirical study, once administered on real captured traffic and analyzed in SPSS/Python, is expected to support.

VI. CONCLUSION

This study proposed Adaptive Hybrid Adversarial Training (AHAT), a framework that unifies multiple adversarial example generators (FGSM, PGD, and Carlini–Wagner) under an epoch-adaptive curriculum scheduler, combined with a hybrid CNN-BiLSTM intrusion detection architecture trained using a composite clean–adversarial–consistency loss. The framework directly addresses two persistent limitations of existing adversarial-training approaches to network intrusion detection: narrow robustness generalization arising from single-technique training, and the erosion of clean-data accuracy that commonly accompanies aggressive fixed-budget adversarial training. The illustrative evaluation pipeline demonstrated that the proposed framework can, in principle, sustain substantially higher mean adversarial accuracy than baseline, single-technique, and non-adaptive ensemble adversarial training approaches, while retaining clean accuracy comparable to an undefended model. Ablation analysis indicated that both attack-family diversity and adaptive curriculum scheduling contribute materially, and largely independently, to these robustness gains. These findings, while illustrative pending full empirical validation, provide a coherent, theoretically grounded, and practically motivated architecture for researchers and practitioners seeking to harden DL-based NIDS against adversarial evasion without compromising day-to-day detection reliability.

VII. RECOMMENDATIONS

Based on the findings and discussion above, the following recommendations are proposed:

1. Security researchers and NIDS developers should adopt multi-generator adversarial training regimes rather than relying on a single attack technique (e.g., PGD alone), given the demonstrated risk of narrow robustness generalization.
2. Curriculum-based, adaptive perturbation scheduling should be preferred over fixed-budget adversarial training, as it appears to materially improve the robustness–accuracy trade-off while easing early-training optimization instability.
3. Institutions intending to operationalize this framework should administer the accompanying data-collection instrument and conduct live model training on captured or benchmark network traffic, followed by SPSS-based statistical validation (ANOVA/Tukey), before the results reported here are treated as conclusive.
4. Security operations centres, particularly in resource-constrained environments, should evaluate lighter-weight architectural variants of the AHAT framework, since the ablation analysis suggests the training regime contributes more to robustness than architectural depth alone.

5. Future research should extend the framework to black-box and transfer-based adversarial attacks, incorporate certified robustness bounds, and evaluate performance on live, continuously evolving traffic rather than static benchmark partitions.
6. Policy-level guidance for national cybersecurity agencies and academic cybersecurity curricula should incorporate adversarial robustness testing as a standard component of NIDS procurement and evaluation, rather than treating it as an optional enhancement.

REFERENCES

[1] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.

[2] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in Proc. 2015 Military Communications and Information Systems Conf. (MilCIS), 2015, pp. 1–6.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. on Learning Representations (ICLR), 2015.

[4] E. Alhajjar, P. Maxwell, and N. Bastian, "Adversarial machine learning in network intrusion detection systems," Expert Systems with Applications, vol. 186, p. 115782, 2021.

[5] K. He, D. D. Kim, and M. R. Asghar, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 538–566, 2023.

[6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. Int. Conf. on Learning Representations (ICLR), 2018.

[7] I. Debicha, T. Debatty, J.-M. Dricot, and W. Mees, "Adversarial training for deep learning-based intrusion detection systems," in Proc. 16th Int. Conf. on Systems (ICONS), 2021, pp. 45–49.

[8] M. M. Rashid, J. Kamruzzaman, M. M. Hassan, T. Imam, S. Wibowo, S. Gordon, and G. Fortino, "Adversarial training for deep learning-based cyberattack detection in IoT-based smart city applications," Computers & Security, vol. 120, p. 102783, 2022.

[9] W. D. Xiong, K. L. Luo, and R. Li, "AIDTF: Adversarial training framework for network intrusion detection," Computers & Security, vol. 128, p. 103141, 2023.

[10] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. 4th Int. Conf. on Information Systems Security and Privacy (ICISSP), 2018, pp. 108–116.

[11] M. M. Hasan, R. Islam, Q. Mamun, M. Z. Islam, and J. Gao, "Adversarial attacks on deep learning-based network intrusion detection systems: A taxonomy and review," SSRN, 2025.

[12] H. Mohammadian, A. A. Ghorbani, and A. H. Lashkari, "A gradient-based approach for adversarial attack on deep learning-based network intrusion detection systems," Applied Soft Computing, vol. 137, p. 110173, 2023.

[13] G. Ghajari, M. K. PK, and F. Amsaad, "Enhancing IoT intrusion detection systems through adversarial training," arXiv preprint arXiv:2507.19739, 2025.

[14] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," IEEE Access, vol. 8, pp. 165130–165150, 2020.

[15] C. Park, J. Lee, Y. Kim, J.-G. Park, H. Kim, and D. Hong, "An enhanced AI-based network intrusion detection system using generative adversarial networks," IEEE Internet of Things Journal, vol. 10, no. 3, pp. 2330–2345, 2022.

[16] Y. Wang et al., "Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey," IEEE Communications Surveys & Tutorials, vol. 25, no. 4, pp. 2245–2298, 2023.

[17] H. Qiu, Y. Zeng, Q. Zheng, S. Guo, T. Zhang, and H. Li, "An efficient preprocessing-based approach to mitigate advanced adversarial attacks," IEEE Transactions on Computers, vol. 73, no. 3, pp. 645–655, 2021.