NOVEMBER 28, 2025

ITERATION 3

CNET

HARRIS HASSAN & ABDULLAH NADEEM & M. MUTTAYAB
DS(D)

Executive Summary

This report documents the final phase of the project, where we moved beyond simple replication of the paper "Enhancing Adversarial Robustness in Network Intrusion Detection" to actively improving its results.

While Iteration 2 successfully replicated the paper's methodology and matched its baseline accuracy (~86.04% Clean / ~83.70% Robust), Iteration 3 introduced significant architectural and training enhancements. The final model (Enhanced ADV_NN) successfully outperforms the paper's proposed model in 8 out of 9 metrics, achieving a new state-of-the-art for this specific experimental setup.

Methodology: From Replication to Enhancement

Feature	**Iteration 2 (Paper Replication)**	**Iteration 3 (Enhanced Model)**	**Impact**
Architecture	Simple Feed-Forward NN (3 Layers)	Deep Residual Network (ResNet)	Allows deeper feature extraction without gradient vanishing, boosting clean accuracy.
Optimizer	Standard Adam	AdamW (Weight Decay)	Better regularization prevents overfitting to the adversarial noise.
Scheduler	None / Static LR	OneCycleLR	"Super-convergence" technique that finds flatter, more stable minima in the loss landscape.
Loss Function	Standard Cross-Entropy	Label Smoothing (0.1)	Prevents the model from being "over-confident," making it harder to fool with small perturbations.
Regularization	smooth=0.5	smooth=1.0	Enforces stricter similarity between clean and attacked feature representations.

Performance Analysis

Clean Data Performance

The most significant improvement in Iteration 3 is the ability to maintain high clean accuracy while being robust. The original paper sacrificed clean accuracy for robustness; Iteration 3 does not.

Metric	Paper / Iteration 2	Iteration 3 (Final)	Improvement
Clean Accuracy	86.04%	87.50%	+1.46%

Robustness Against PGD Attack (The "Gold Standard")

Epsilon (ϵ)	Paper / Iteration 2	Iteration 3 (Final)	Improvement	Status
0.01	85.90%	87.11%	+1.21%	✓ Superior
0.05	85.13%	85.83%	+0.70%	✓ Superior
0.10	83.70%	84.93%	+1.23%	✓ Superior
0.15	83.70%	83.80%	+0.10%	✓ Superior

Robustness Against FGSM Attack

FGSM is a simpler, single-step attack.

Epsilon (ϵ)	Paper / Iteration 2	Iteration 3 (Final)	Improvement	Status
0.01	85.90%	87.13%	+1.23%	✓ Superior
0.05	85.15%	86.03%	+0.88%	✓ Superior
0.10	83.77%	84.48%	+0.71%	✓ Superior
0.15	80.49%	80.32%	-0.17%	⚠ Comparable

Note: The minor dip (-0.17%) in FGSM at epsilon=0.15 is a negligible trade-off for the massive gains (+1.46%) in Clean Accuracy and PGD robustness.

Efficiency & Lightweight Analysis

1st Model Complexity:

- Iteration 2 uses a simple Feed-Forward Neural Network (FFNN) with only 2 hidden layers (128 \to 64 neurons). This requires very few matrix multiplications (FLOPs) per sample.
- Iteration 3 uses a ResNet Architecture. It includes input projection layers, multiple residual blocks with batch normalization, and skip connections. This significantly increases the parameter count and computational depth.

2nd Training Cost:

- Iteration 2 trains for fewer epochs (20) with a simpler backward pass.
- Iteration 3 was increased to 25 epochs and uses 7 iterations for PGD generation during training (vs 5 or fewer in lighter setups). This makes the training loop roughly 40-50% slower than Iteration 2.

Conclusion on Efficiency:

For any modern server or standard edge device, the new model is still lightweight enough to run in real-time while providing significantly better security coverage.