ICS Cyber-Threat Detection: Towards an Effective Adversarial Sample Generation and DNN Defence

Urslla Uchechi Izuazu, Vivian Ukamaka Ihekoronye, Dong-Seong Kim, Jae Min Lee
Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, South Korea
uursla8@gmail.com, ihekoronyevivian@gmail.com, (dskim, ljmpaul)@kumoh.ac.kr

***Abstract*—This study introduces CIGM, a novel adversarial sampling method suitable for ICS network environments, aiming to simulate adversary attacks to fool deep learning (DL)intrusion detection systems(IDM) towards misclassification. Leveraging adversarial training as a defense mechanism, the proposed DL-based IDS exhibits resilience with CIGM, compared to existing techniques, highlighting its effectiveness for optimal threat detection within the ICS network environment.**

***Index Terms*—Intrusion Detection, Adversarial Attack, Industrial control system, DNN**

I. Introduction

The integration of IoT with industrial control systems (ICS) enhances operational excellence but exposes these systems to cyber threats [1], [2]. While deep neural networks (DNN) have shown improved detection capabilities for attack detection within ICS networks, their susceptibility to adversarial attacks remains a significant concern. Adversaries exploit vulnerabilities in DNN by introducing carefully crafted perturbations to input data, leading to misclassifications and undermining the reliability of the IDS. Addressing this vulnerability is essential for enhancing the robustness of DNN in detecting and mitigating cyber threats in ICS networks.

Recent studies have advanced ICS security by adopting various adversarial methods and defense mechanisms to enhance resilience against perturbations to DNN models. In [3], scholars evaluated the performance of two DL models;feed-forward neural network (FNN) and self-normalizing neural network (SNN), and also investigated the effects of adversarial attacks on their proposed model using the fast gradient sign method (FGSM) and basic iteration method (BIM) methods. FGSM takes a single step, while BIM iteratively adjusts input data.

However, these methods (FGSM /BIM) involve minor adjustments to every sample feature utilizing a fraction of the model's gradient sign. Although gradient-based methods excel in computer vision, where subtle pixel alterations go unnoticed by humans, their direct application to ICS is problematic. Uncontrolled feature modifications may render samples unprocessable or incomprehensible to ICS devices, thwarting attacks from reaching their designated targets and reducing their impact on the physical environment. Thus, **this study introduces a novel adversarial sampling technique, the Controlled Iterative Gradient Method (CIGM),** which addresses the shortcomings of existing attacks by precisely controlling feature modifications. CIGM is tailored for real-world ICS network environments and relevant to broader cybersecurity applications.

II. Methodology

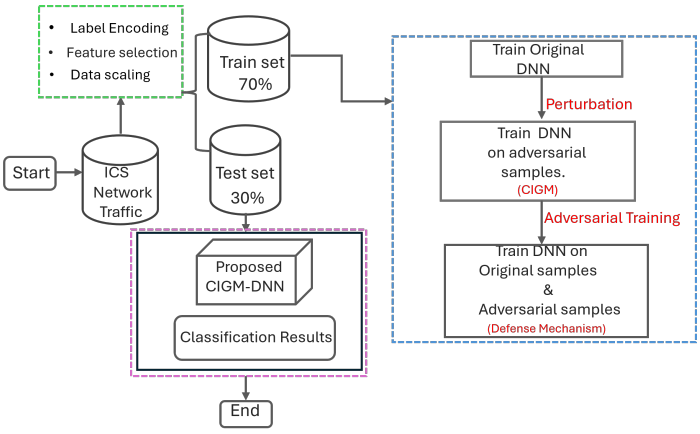

Fig. 1. Proposed CIGM-DNN WorkFlow

The workflow shown in Fig. 1 outlines the process used in designing the CIGM-DNN model. It commenced with importing and preprocessing the ICS network traffic, which is then split into a training set (70%) and a test set (30%). The original DNN is first trained on the original samples. Furthermore, the CIGM is applied to generate adversarial samples, introducing perturbations to the DNN model. Consequently, to protect the DNN model from adversarial perturbation, adversarial training as a defense mechanism was employed. Finally, the proposed CIGM-DNN model is evaluated based on key evaluation metrics

A. Description of Dataset/ Preprocessing

The WUSTL-IIOT-2021 dataset is derived from a real ICS network testbed, specifically monitoring water levels and turbidity in water storage tanks, including various attack types and normal traffic [4].

B. Experimental Setup and Hyperparameters

Experiments were conducted using TensorFlow 2.9.0 in a Python environment on a system with an Intel(R) Core(TM) i7-7400 CPU @ 3.00GHz, 16GB RAM, and a Tesla K80 GPU. Hyperparameters were manually adjusted for optimal performance, as summarized in Table I.

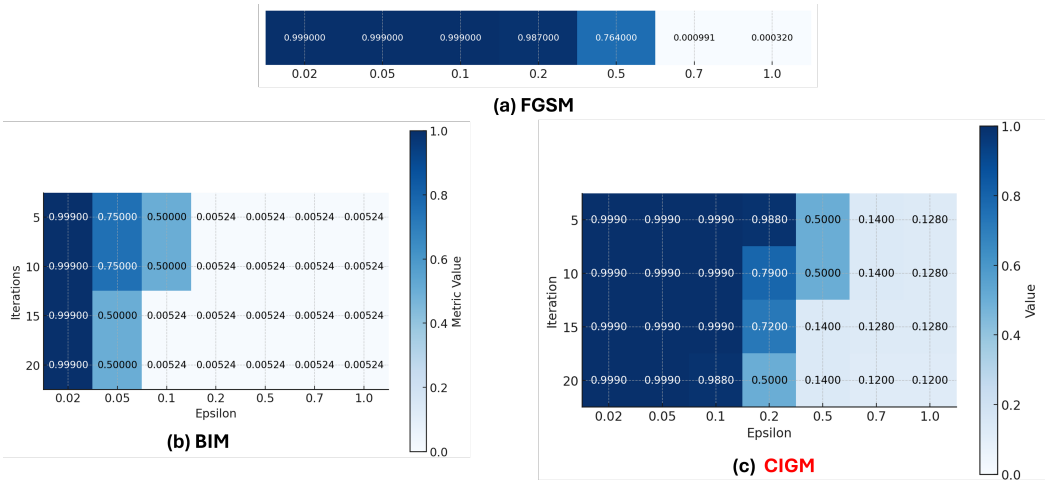

Fig. 2. Accuracy Decline Comparison of Proposed CIGM and Existing Methods with Increase in the Number of Epsilon and Iterations

TABLE I
HYPERPARAMETERS EMPLOYED

Hyperparameter	**Value**
Number of Layers	5
Optimizer	Relu/Softmax
Batch Size	30
Activation Function	Adam
Epochs	20
Learning Rate	0.001
Loss Function	Sparse Categorical Cross-Entropy
Epsilon	0.1
iteration	5

III. PERFORMANCE EVALUATION/RESULT ANALYSIS

To show the superiority of our proposed CIGM over FGSM and BIM, we analyzed accuracy performance based on the stable accuracy of 0.999000. Fig. 2 shows the decline in accuracy across the different adversarial methods. In Fig.2 (a), FGSM exhibits a noticeable decline in accuracy at epsilon 0.5, reaching 0.764000. Similarly, Fig. 2(b) shows accuracy trends with the BIM method, with significant decreases observed at epsilon 0.1. Fig. 2(c) displays the accuracy decline using our CIGM, which maintains higher accuracy even at higher epsilon values and iterations compared to FGSM and BIM. Our method demonstrates the lowest accuracy reduction, highlighting its robustness and resilience against adversarial attacks.

A. Proposed CIGM-DNN in Comparison with State-of-the-Art

Our proposed model surpassed existing DL-based approaches, achieving a commendable accuracy of 99.90% as table II indicates, evidencing its superior predictive performance.

IV. CONCLUSION

This study introduced a novel and effective adversarial sampling approach, suitable for real-world ICS networks environment. The implementation of adversarial training as a defense mechanism enhances overall model performance. Future work aims at securing proposed CIGM DNN in blockchain to avoid unauthorized access or tempering from adversaries.

TABLE II
COMPARISON OF ADVERSARIAL RESILIENCE OF PROPOSED CIGM DNN WITH EXISTING DL MODELS

Ref.	Model	ACC
[3]	SNN	0.9821
[3]	FNN	0.9595
Ours	**CIGM-DNN**	**0.9990**

ACKNOWLEDGMENT

This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of IITP grant funded by the Korea government(MSIT) (IITP-2024-2020-0-01612, 50%) and by Priority Research Centers Program through the NRF funded by the MEST(2018R1A6A1A03024003, 50%)

REFERENCES

[1] U. U. Izuazu, V. U. Ihekoronye, D.-S. Kim, and J. M. Lee, "Securing critical infrastructure: A denoising data-driven approach for intrusion detection in ics network," in *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 2024, pp. 841–846.

[2] L. A. C. Ahakonye, C. I. Nwakanma, J.-M. Lee, and D.-S. Kim, "Scada intrusion detection scheme exploiting the fusion of modified decision tree and chi-square feature selection," *Internet of Things*, vol. 21, p. 100676, 2023.

[3] O. Ibitoye, O. Shafiq, and A. Matrawy, "Analyzing adversarial attacks against deep learning for intrusion detection in iot networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[4] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, "Machine learning-based network vulnerability analysis of industrial internet of things," *IEEE internet of things journal*, vol. 6, no. 4, pp. 6822–6834, 2019.