Adversarial Attack of ML-based Intrusion Detection System on In-vehicle System using GAN

EunSeong Seo
Department of Automobile Convergence
Korea University
Seoul, Korea
seoun4612@korea.ac.kr

JeongEun Kim
Department of Automobile Convergence
Korea University
Seoul, Korea
mine1372@korea.ac.kr

Wook Lee
School of Electrical Engineering
Korea University
Seoul, Korea
leewook94@korea.ac.kr

Junhee Seok*
School of Electrical Engineering
Korea University
Seoul, Korea
jseok14@korea.ac.kr

***Abstract*—** In recent years, research has focused on developing intrusion detection systems (IDS) within vehicle networks to prevent automotive hacking from external cyberattacks. While machine learning (ML) techniques have shown promise in detecting known attacks, their vulnerability to adversarial examples remains a significant challenge. In this study, we introduce a Generative Adversarial Network (GAN)-based method for creating adversarial attacks capable of bypassing ML-based IDS in in-vehicle networks. Our approach involves preprocessing an automotive hacking dataset, training a GAN-based model, and evaluating the generated attacks using accuracy metrics. The results demonstrate that adversarial attacks effectively reduce the detection accuracy of various IDSs to less than 50%, emphasizing the importance of addressing adversarial cases when designing and evaluating ML-based IDSs for in-vehicle networks. Additionally, t-SNE visualization reveals the successful generation of new adversarial attacks, highlighting the need for ongoing research to strengthen the security of in-vehicle systems.

Keywords—GAN, Adversarial attack, Machine Learning, In-vehicle networks, Intrusion detection system

I. Introduction

The advancement of deep learning techniques has fostered research across a multitude of domains, including vehicles, weather, and financial data [1-3]. The vehicular domain, in particular, has greatly benefited from the integration of deep learning algorithms, resulting in enhanced safety and efficiency. Nevertheless, to effectively utilize deep learning algorithms, it is imperative to conduct research on preprocessing non-standardized big data [4-5]. Moreover, the need to distinguish data with malicious intent is emerging as a critical challenge, especially within vehicular networks. Vehicles employ the Controller Area Network (CAN) communication protocol to facilitate efficient interactions between electronic control units (ECUs) and sensors. However, the CAN bus's utilization of an uncertified broadcasting communication method exposes vehicles to significant safety risks through external communication devices such as Wi-Fi or Bluetooth. To tackle these issues, researchers have developed machine learning (ML)-based intrusion detection systems for vehicular networks [6]. Despite the promising performance of ML-based intrusion detection systems in identifying known attacks, their susceptibility to adversarial examples presents a considerable challenge. As a result, devising adversarial attacks capable of bypassing these ML-based intrusion detection systems is critical for uncovering potential vulnerabilities and encouraging the development of more robust countermeasures. Owing to the limited availability of real-world big data, numerous simulation-based studies have been conducted [7,8]. In this paper, we introduce a novel method for generating adversarial attacks that can circumvent ML-based intrusion detection systems, underscoring the necessity for improved security measures in vehicular networks.

II. Background

A. Controller Area Network data

CAN communication works by exchanging messages consisting of a unique identifier (ID), data length code (DLC) and data fields. Each message is transmitted over the CAN bus and the origin of the data is unknown. However, this CAN bus-based process can be exploited by attackers to launch various attacks, such as fuzzy attacks. In a fuzzy attack, an attacker can change the data fields being transmitted, potentially causing abnormal behavior or damage to vehicle systems.[9].

B. Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of machine learning models consisting of generators and discriminators that operate on competition.[10] The Generator generates synthetic data samples, and the Discriminator evaluates the authenticity of real and generated samples. GANs have been used in a variety of applications.[11] In the context of

intrusion detection systems, GANs have been used to generate adversarial examples, as demonstrated in IDSGANs.[12].

C. Adversrial Attack

Adversarial attacks are designed to trick machine learning models by introducing small perturbations in the input data, causing the model to misclassify or misinterpret the data.[13] These attacks can severely impact the performance of ML-based intrusion detection systems and expose unique vulnerabilities. By studying adversarial attacks, researchers can develop more robust models and countermeasures against potential threats.

III. METHOD

A. Data collection and preprocessing for IDS learning

In this study, the Fuzzy attack dataset of Hyundai YF Sonata of "In-vehicle Network Intrusion Detection track" of 'Information Security R&D Data Challenge 2019' developed by Hacking and Countermeasure Research Lab, Korea was used to evaluate the model. [14] As shown in Table 1, the data set contains normal data and about 21% attack data. The preprocessing step converted the hexadecimal data to decimal and normalized the data by dividing the ID, DLC, and data field values by their respective maximum values (4096, 8, and 256). This normalization ensures that the data are of similar size and suitable for machine learning models. Also, we labeled normal data as 0 and attack data as 1.

TABLE I. NORMAL DATA AND ATTACK DATA USED FOR INTRUSION DETECTION SYSTEM LEARNING

Dataset	Train/Test	Total data	Normal	Attack
Hyundai YF Sonata	Train	403,299	318,655	84,644 (20.99%)
	Test	351,273	276,337	74,936 (21.33%)

B. Model Architecture

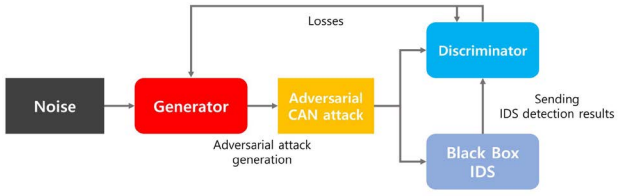

Fig. 1. proposed model architecture.

The proposed model architecture is based on IDSGAN and consists of two main components: generator and discriminator. The generator creates an adversarial attack from random noise and injects it into a discriminator and an intrusion detection system (IDS), and the discriminator mimics the IDS. The overall structure is shown in Figure 1. The purpose of generators is to generate attacks that IDS does not detect, while the purpose of discriminators is to accurately distinguish between real and generated attacks. This model is trained using the loss function of Equation 1 for the generator and Equation 2 for the discriminator to optimize the performance of the two components.

$$L_G = \mathbb{E}_{M \in S_{attack}, N} D(G(M, n)) \quad (1)$$

$$L_D = \mathbb{E}_{s \in B_{normal}} D(S) - \mathbb{E}_{s \in B_{attack}} D(S) \quad (2)$$

IV. EXPERIMENTS RESULTS

A. Setting

Experiments involved training the proposed GAN-based model using a preprocessed fuzzy attack dataset. The trained model was used to generate adversarial attacks injected into several ML-based intrusion detection systems to evaluate their robustness against these attacks. The intrusion detection system included machine learning models of multi-layer perceptron (MLP), random forests (RF), and logistic regression (LR) to evaluate the generalizability of adversarial attacks. In the first experiment, normal CAN data and fuzzy attack data were injected at a ratio of 50:50, and in the second experiment, normal CAN data and generated adversarial attack data were injected at a ratio of 50:50.

B. Evaluation

Accuracy was used in Equation (3) to evaluate the performance of the generated adversarial attack. Accuracy measures the ratio of true normal to true abnormal. The lower the accuracy of the IDS, the less robust the IDS is.

$$accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (3)$$

C. Results

Injecting the generated adversarial attacks into a selected ML-based intrusion detection system reduced detection accuracy to less than 50% for all models. This result suggests that the adversarial attack generated by our GAN-based method was effective in bypassing the IDS, highlighting the potential weakness of current ML-based intrusion detection techniques.

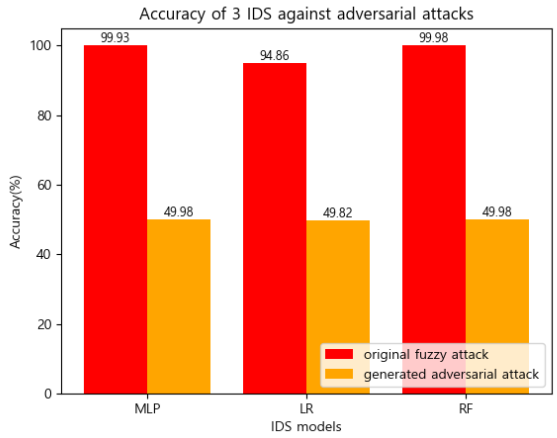

Fig. 2. Accuracy of 3 IDS against fuzzy attack and adversarial attack

To gain further insight into the generated adversarial attack, we use t-distributed stochastic neighbor embedding (t-SNE) to obtain 1000 random normal data, 1000 fuzzy attack data and 1000 adversarial attack data each. Sampling was performed to visualize the data distribution. Adversarial attacks generated through visualization are far from both normal and real attacks. These results suggest that our method succeeds in creating a new adversarial case for bypassing IDS and requires further investigation to strengthen the security of in-vehicle networks.

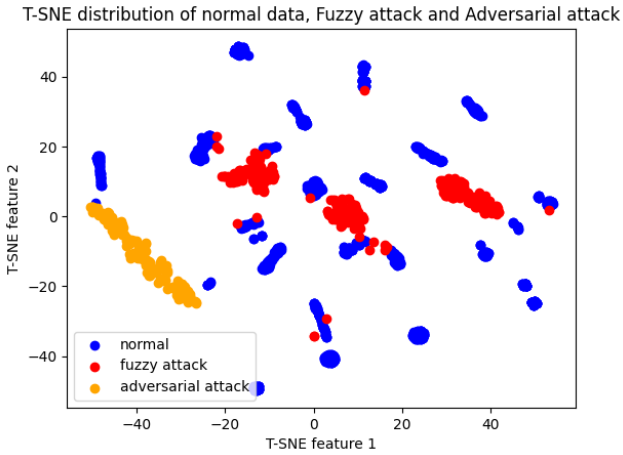

Fig. 3. t-SNE distribution of normal data, Fuzzy attack and Adversarial attack

V. Conclusion

In this study, we introduced a GAN-based method for generating adversarial attacks for ML-based intrusion detection systems in in-vehicle networks. By exploiting vulnerabilities in these systems, our method exposes potential security risks and demonstrates the need for more robust intrusion detection techniques. These results highlight the importance of considering adversarial attacks when designing and evaluating ML-based intrusion detection systems and the need for continued research in this area to ensure the safety and security of in-vehicle systems.

Acknowledgment

This work was supported by a grant from the National Research Foundation of Korea (NRF-2022R1A2C2004003)

References

[1] Kuutti, Sampo, et al. "A survey of deep learning applications to autonomous vehicle control." *IEEE Transactions on Intelligent Transportation Systems* 22.2 (2020): 712-733.

[2] Ahn, Heewoong, et al. "Searching similar weather maps using convolutional autoencoder and satellite images." *ICT Express* 9.1 (2023): 69-75.

[3] Ozbayoglu, Ahmet Murat, Mehmet Ugur Gudelek, and Omer Berat Sezer. "Deep learning for financial applications: A survey." *Applied Soft Computing* 93 (2020): 106384.

[4] Al-Shemarry, Meeras Salman, and Yan Li. "Developing learning-based preprocessing methods for detecting complicated vehicle licence plates." *IEEE Access* 8 (2020): 170951-170966.

[5] Choi, Sunho, et al. "Conversion of Automated 12-Lead Electrocardiogram Interpretations to OMOP CDM Vocabulary." *Applied Clinical Informatics* 13.04 (2022): 880-890.

[6] Rajapaksha, Sampath, et al. "AI-Based Intrusion Detection Systems for In-Vehicle Networks: A Survey." *ACM Computing Surveys* 55.11 (2023): 1-40.

[7] Seo, Jangwon, Insoo Kim, and Junhee Seok. "Grid-wise simulation acceleration of the electromagnetic fields of 2D optical devices using super-resolution." *Scientific Reports* 13.1 (2023): 435.

[8] Hwang, Hyo-Seok, Minhyeok Lee, and Junhee Seok. "Deep reinforcement learning with a critic-value-based branch tree for the inverse design of two-dimensional optical devices." *Applied Soft Computing* 127 (2022): 109386.

[9] Han, Mee Lan, Byung Il Kwak, and Huy Kang Kim. "Event-triggered interval-based anomaly detection and attack identification methods for an in-vehicle network." *IEEE Transactions on Information Forensics and Security* 16 (2021): 2941-2956.

[10] Goodfellow, Ian, et al. "Generative adversarial networks." *Communications of the ACM* 63.11 (2020): 139-144.

[11] Kim, Insoo, Minhyeok Lee, and Junhee Seok. "ICEGAN: inverse covariance estimating generative adversarial network." *Machine Learning: Science and Technology* 4.2 (2023): 025008.
[12] Lin, Zilong, Yong Shi, and Zhi Xue. "Idsgan: Generative adversarial networks for attack generation against intrusion detection." *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part III*
[13] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
[14] Han, Mee Lan, Byung Il Kwak, and Huy Kang Kim. "Anomaly intrusion detection method for vehicular networks based on survival analysis." *Vehicular communications* 14 (2018): 52-63.