Adversarial Attacks in Intrusion Detection Systems: Triggering False Alarms in Connected and Autonomous Vehicles

Fatimah Aloraini*
School of Computer Science and
Informatics, Cardiff University,
Cardiff, United Kingdom
Email: alorainif@cardiff.ac.uk

Amir Javed
School of Computer Science and
Informatics, Cardiff University,
Cardiff, United Kingdom
Email: javeda7@cardiff.ac.uk

***Abstract*—As connected and autonomous vehicles (CAVs) proliferate, securing their internal vehicle networks (IVNs) against cyber threats is paramount. Current research focuses on developing intrusion detection systems (IDSs) using machine learning (ML) models to handle diverse threats. However, ML-based IDSs introduce significant risks from adversarial attacks. This paper investigates the vulnerability of ML-based IDSs in IVNs to such attacks. It shifts focus from manipulating malicious frames to appear benign to exploring IDS susceptibility to benign frames appearing malicious, potentially triggering false alarms. In critical safety applications like CAVs, these alarms can compromise safety and operational integrity. We studied IVN traffic and designed adversarial samples simulating potential threats. Our experiments, using five ML algorithms and four state-of-the-art adversarial methods, demonstrate an attack success rate of up to 89%. This underscores the urgent necessity to address this vulnerability, as neglecting it renders IDSs ineffective and increases the risk of vehicle manipulation.**

***Keywords*—adversarial machine learning; connected and autonomous vehicle; controller area network; in-vehicle network; cybersecurity**

I. Introduction

As we advance toward the future, connected and autonomous vehicles (CAVs) are becoming integral to transportation systems. CAVs are employed in various fields such as road safety, traffic management, and data-driven mobility, offering new business opportunities across multiple industries including transportation, retail, finance, insurance, energy, health services, and media [1]. This expansive application leads to significant market potential, projected to reach $7 trillion by 2050 [2]. As the market for CAVs grows, it creates new cybersecurity vulnerabilities with severe implications for CAV safety.

The vehicle system consists of a complex cyber-physical network. Electronic control units (ECUs) within the vehicle communicate through the internal vehicle network (IVN), primarily using the controller area network (CAN) protocol, the de facto standard for IVNs [3]. Designed in the 1980s, the CAN protocol emphasized reliability, cost-effectiveness, and a bus topology that ensures high-integrity real-time communications, assuming an isolated environment where security was not a concern [4]. Consequently, the CAN protocol has inherent vulnerabilities, including the absence of authentication mechanisms, lack of encryption, the broadcast nature of transmission, and an identifier-based priority scheme, which facilitates denial-of-service (DoS) attacks by injecting high-priority identifiers [5], [6]. Researchers have demonstrated successful remote access to CAN-based vehicles such as the Jeep Cherokee [7], Tesla [8], and BMW [9]. For example, Miller and Valasek [7] demonstrated a successful hack of a Jeep Cherokee, controlling it remotely via the Internet using a laptop.

Therefore, considerable efforts have been made to protect vehicles from security threats. As a reactive security mechanism, current research has focused on developing intrusion detection systems (IDSs) for IVNs. IDSs can be categorized as either signature-based or anomaly-based. Anomaly-based detection approaches, which rely on machine learning (ML) and deep learning (DL), have garnered attention due to their capability to detect novel attacks—a limitation of signature-based IDSs [10], [11].

However, integrating ML/DL into CAVs introduces substantial cybersecurity concerns. Studies, including [12], [13], have exposed the vulnerabilities of ML/DL models to a unique category of threats known as 'adversarial attacks.' These techniques manipulate input data, causing ML models to misclassify and respond inappropriately. Deploying IDSs without considering their susceptibility to adversarial manipulation not only fails to protect but potentially escalates the risk of vehicle manipulation.

Thus, research has shifted towards understanding the adversarial manipulation of IDS solutions. However, previous studies have primarily concentrated on the manipulation of malicious samples to appear normal and bypass the IDS, which is an expected scenario from the adversary. To the best of our knowledge, no studies have evaluated the robustness of ML-based IDS by manipulating benign traffic to appear as various attacks, such as fuzzy and spoofing attacks. Such false alarms in safety-critical applications like vehicles could lead to inappropriate responses, resulting in life-threatening situations, financial damage, and potential legal liabilities for manufacturers. These scenarios can cause harm without an actual attack payload; they only require manipulating benign frames to fool the IDS into raising a false alarm that triggers an inappropriate vehicle response. The main contributions of this paper are as follows:

- Introduction of a novel adversarial strategy designed to manipulate benign frames within IVNs against IDS, illustrating the potential for false alarms to trigger unintended defensive responses.
- Identification of potent adversarial techniques for crafting adversarial samples from benign IVN frames, demonstrating their capacity to undermine IDSs deployed in IVNs, thus amplifying safety and security concerns.

II. Related Work

In exploring adversarial manipulations of IDS, the literature predominantly focuses on techniques to craft adversarial samples misclassified by IDS. These techniques are categorized into gradient-based, evolutionary, and generative adversarial network (GAN)-based approaches. A summary of the related work is provided in Table I.

Gradient-based methods are notable for generating adversarial samples against ML/DL-based IDSs.Researchers like Papadopoulos et al. [14], Guo et al. [15], and Pacheco et al. [16] demonstrated that techniques such as the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Jacobian-based Saliency Map Attack (JSMA), and Carlini & Wagner Attack (C&W) reduced the efficacy of models like Support Vector Machine (SVM), Decision Tree (DT), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), k-nearest Neighbors (kNN), Multilayer Perceptron (MLP), and Residual Network (ResNet). Owezarski [17] showed that statistical perturbations targeted ML-based IDS systems, particularly Random Forest (RF), affecting models like SVM, kNN, and Long Short-Term Memory (LSTM).

GAN-based approaches for crafting adversarial samples have been explored by Alhajjar et al. [18] and Han et al. [19]. Alhajjar et al. studied feature-level attacks on eleven ML models and a voting classifier, while Han et al. conducted traffic-level attacks, achieving high evasion rates on Kitsune and other ML-based IDSs. Pillai et al. [21] revealed vulnerabilities in GAN-trained IDS using FGSM. Shu et al. [22] combined active learning with GANs but overlooked domain constraints, leading to non-functional traffic flows. Lin et al. [23] introduced IDSGAN, which modifies unimportant traffic features and combines them with the original important features to evade detection while maintaining attack functionality. Usama et al. [24] proposed a similar method, enhancing robustness via adversarial training. Duy et al. [25] explored Wasserstein GAN with Gradient Penalty (WGAN-GP), WGAN-GP with the two timescale update rule (WGAN-GP TTUR), and adversarial generative adversarial network (AdvGAN) to generate perturbed samples, effectively deceiving IDSs in software-defined networking environments.

TABLE I
Summary of Related Work on Adversarial Attacks Against IDSs

Ref.	Adversarial Method	Target Model	Manipulation Category
[17]	Statistical perturbation	RF, SVM, kNN, LSTM	FN
[14]	FGSM	SVM, ANN	FN
[15]	BIM	CNN, SVM, kNN, MLP, ResNet	FN
[16]	JSMA, FGSM, C&W	MLP, DT, RF, SVM	FN
[18]	Evolutionary methods, GANs	ML models, voting classifier	FN
[19]	Evolutionary methods, GANs	ML models, Kitsune [20]	FN
[21]	FGSM	GAN	FN
[22]	GAN	Gradient Boosted DT	FN
[23]	GAN	NB, LR, SVM, MLP, DT, RF, kNN	FN
[24]	GAN	NB, LR, SVM, DNN, DT, RF, kNN, GB	FN
[25]	WGAN	DT, LR, CNN, MLP, LSTM	FN
Our work	FGSM, BIM, PGD, DT	DNN, DT, RF, ET, XGBoost	FP

While previous research has demonstrated the efficacy of adversarial attacks in deceiving IDSs by transforming malicious scenarios into seemingly normal ones—thereby bypassing detection and creating false negatives (FN)—there exists a significant gap in studying the inverse scenario. The possibility of normal activities being misclassified as malicious results in false positives (FP). These FPs are not mere inconveniences but potential hazards that compromise the safety and operational integrity of CAVs. Therefore, it is imperative to explore this aspect of adversarial attacks to ensure IDSs accurately distinguish between threats and legitimate activities, thereby maintaining the safety of CAVs.

III. Methodology

A. Threat Model

We investigate a scenario where an adversary induces misclassification in ML/DL-based IDS within an IVN by crafting adversarial samples from CAN traffic. The adversary can create these samples in advance and inject them in real-time by exploiting CAN bus vulnerabilities, as discussed in Section I. To facilitate future comparisons, our threat model is defined according to the taxonomy dimensions outlined by Huang et al. [12], as summarized in [26]. These dimensions include the influence, specificity, impact, knowledge, and goal of the adversarial attack.

Fig. 1. The Standard Structure of CAN Data Frame

Our proposed attacks occur during the inference phase, after the IDS model has been trained and tested. The adversary targets benign CAN frames, manipulating them to generate false alarms and trigger misclassification. We assume a powerful, white-box adversary with complete system knowledge to thoroughly understand IDS vulnerabilities in CAVs. This understanding is crucial for designing effective defense and response strategies. Additionally, since adversarial attacks are transferable, successful white-box attacks suggest potential applicability to other black-box IDS models trained in the same context [27], [28]. The adversary's goal is to compromise the integrity of the IDS, thereby undermining user trust in its defense mechanism.

B. IDS Architecture

The CAN frame transmitted over the IVN is inherently simple [6], focusing on a limited set of features: the identifier (ID), data length code (DLC), data field (up to eight bytes), cyclic redundancy check (CRC), acknowledgment (ACK), and bits for the start (SoF) and end (EoF) of the frame. Figure 1 shows the structure of a standard CAN data frame. IVN-based IDSs typically use features like the ID alone, the payload alone, or a combination of ID and payload. Our IDS uses the entire CAN frame, detecting changes in both IDs and payloads. According to Rajapaksha et al. [10], this comprehensive method is the most extensively studied in the literature, with the highest number of publications compared to other IDS types.

We implemented five IDSs using a combination of DL and ML algorithms: DNN, DT, RF, Extra Trees (ET), and XGBoost. For the DL component, we chose a DNN due to its simplicity and practicality in resource-constrained environments like vehicles, compared to more complex spatial and sequential models such as CNNs. DNNs also enable the use of the entire CAN frame features, unlike CNNs, which typically focus on sequences identified by CAN IDs and may miss payload attacks. Regarding the ML algorithms, based on a recent survey [29] that showed ML algorithms used as IVN-based IDS, we chose tree-based models due to their proven effectiveness. According to [17], [30], these models, including DT, RF, ET, and XGBoost, handle nonlinear network traffic data effectively and often outperform other ML algorithms like NB and kNN on complex datasets. Additionally, tree-based models calculate feature importance during training, aiding in feature engineering, and their inherent randomness enhances robustness and generalizability.

The DNN model employs an architecture with 10 neurons in the input layer and 5 neurons in the output layer. It comprises 4 hidden layers, each containing 16 neurons. The model is trained over 50 epochs with a batch size of 32. The *ReLU* activation function is utilized for the hidden layers, while the *softmax* function is applied in the output layer. The *Adam* optimizer is employed, and the loss function is *categorical cross-entropy*. The other ML-based models were built using the default scikit-learn and XGBoost configurations. This set of IDSs will serve as the target for our proposed adversarial attack scenario.

TABLE II
STATISTICAL BREAKDOWN OF THE CAR HACKING DATASET [31]

Attack Type	Benign	Malicious	Total
DoS	3,078,250	587,521	3,665,771
Fuzzy	3,347,013	491,847	3,838,860
Gear Spoofing	3,845,890	597,252	4,443,142
RPM Spoofing	3,966,805	654,897	4,621,702
Total	14,237,958	2,331,517	16,569,475

C. Dataset

The primary dataset used in this work is the car hacking dataset [31]. It was chosen due to our objective of assessing the vulnerabilities of IVN-based IDSs against adversarial manipulations that could lead to FPs. As noted in a recent survey [10], this dataset is the most frequently used for IVN-based IDS development, making it representative of the state-of-the-art. Additionally, it is based on real traffic data from an actual vehicle, providing realistic conditions. The dataset comprises five segments: one for normal driving data and four for different IVN attacks, including DoS, fuzzing, and two types of spoofing attacks on RPM and gear displays.Each segment spans 30 to 40 minutes and includes attributes like timestamp, CAN ID, DLC, payload, and labels distinguishing normal from malicious frames. Refer to Table II for a detailed statistical breakdown.

To prepare the dataset for IDS training and testing, the following preprocessing steps were applied:

- Adjusting label misplacement: The original dataset has 12 columns: Timestamp, CAN ID, DLC, eight data fields (D0 to D7), and a label column. Labels were misplaced into the first null data field (e.g., D4) when the DLC was less than 8. An automated Python script was developed to correct this by repositioning the labels into the correct column.
- Merging subsets: The dataset was divided into folders for each attack type, with each folder containing both normal and attack traffic. These subsets were then merged into a comprehensive dataset to train a single IDS capable of identifying all attack types.
- Feature selection: All features except the timestamp were used, as timestamps are generally disregarded in the literature unless explicitly required by the detection mechanism.
- Hexadecimal to decimal conversion: The CAN ID and data fields (D0 to D7) were logged as hexadecimal values and converted to decimal format for compatibility with ML/DL algorithms.

TABLE III
PERFORMANCE METRICS OF THE TARGET IDSS ON THE TEST SET UNDER BENIGN SETTINGS

Model	Benign Samples	Malicious Samples	F1 score	FN	FP
DNN	4,272,006	698,837	99%	754	243
DT			100%	23	1
RF			100%	0	0
ET			100%	0	0
XGBoost			100%	4	0

Following the preprocessing steps, the IDS models were trained on a merged dataset comprising four subsets, totaling 16,569,475 samples. This dataset includes 14,237,958 normal and 2,331,517 malicious samples, split into 70% for training (11,598,632 samples) and 30% for testing (4,970,843 samples). The IDS classifies input frames into five categories: normal, DoS, fuzzy, gear spoofing, or RPM spoofing. Given the dataset's imbalance, the IDS's performance was evaluated using the F1-score metric, with the results presented in Table III.

D. Domain Constraints

With the IDS models trained and tested, we proceeded to generate adversarial samples. The initial step involved understanding domain constraints to ensure that adversarial frames are valid CAN frames. Mbow et al. [32] highlight that few studies consider these constraints when developing adversarial attacks on network traffic, noting that methods effective in other contexts may not perform well in network environments. The configurable features of a CAN data frame include the ID, DLC, and data fields. Each ECU responds only to a predefined set of IDs specific to its functions; thus, maintaining the ID feature is crucial for the proper processing of adversarial frames. The DLC indicates the number of bytes in the data field, which must be between 1 and 8 as specified by the CAN protocol [3]. Since the DLC is configured during setup, it cannot be manipulated if adversaries are manipulating existing frames logged from CAN traffic. Only the data field, consisting of 8 dynamic bytes, can be realistically manipulated by adversaries.

To enforce these constraints, we used a boolean mask during adversarial sample generation, marking only the data fields as "True" for manipulation. We also applied clipping values to keep manipulations within the 0 to 255 range. Compliance with these constraints was confirmed through a post-generation check. Table IV summarizes the applied constraints.

E. Adversarial Attack Method

The adversarial attack problem is formulated as an optimization task, aiming to identify the minimal perturbation (epsilon) that causes the target IDS to misclassify an input. Our goal is to evaluate the vulnerability of the IVN-based IDS to manipulated normal frames that can raise false alarms. To achieve this, we utilize two main methodologies: gradient-based attacks and DT attacks, assuming a powerful adversary with access to both the dataset and IDS models.This approach allows us to gain a realistic understanding of the IVN-based IDS vulnerabilities. Several techniques fall under the gradient-based category for generating adversarial attacks. These techniques vary in their approaches to calculating epsilon, influencing the speed and strength of the generated adversarial examples. Based on recent work [33], we selected the following techniques, which demonstrated their effectiveness against IVN-based IDSs:

- FGSM [34]: generates adversarial examples quickly by using the gradient of the loss with respect to the input data.
- BIM [35]: iteratively applies FGSM to create stronger adversarial examples, trading off some speed for improved attack strength.
- Projected Gradient Descent (PGD) [36]: extends BIM by projecting the perturbations back onto an epsilon ball, ensuring the perturbations stay within a defined limit.
- DT Attack [37]: identifies paths to leaf nodes with different class labels and modifies specific features to induce misclassification in tree-based models.

We applied these techniques in a novel way by starting with a normal sample and adding perturbations under IVN constraints to misclassify it as an attack, generating FP alarms that could harm the vehicle. As depicted in Figure 2, once the IDS models were trained and tested, we extracted all 14,237,958 normal frames from the dataset. These frames,

TABLE IV
CONSTRAINTS APPLIED IN THE GENERATION OF ADVERSARIAL EXAMPLES FOR CAR HACKING DATASET [31]

CAN Field	Range	Modification	Mask
CAN ID	[0, 1068]	No	False
DLC	[1, 8]	No	False
D0	[0, 255]	Yes	True
D1	[0, 255]	Yes	True
D2	[0, 255]	Yes	True
D3	[0, 255]	Yes	True
D4	[0, 255]	Yes	True
D5	[0, 255]	Yes	True
D6	[0, 255]	Yes	True
D7	[0, 255]	Yes	True

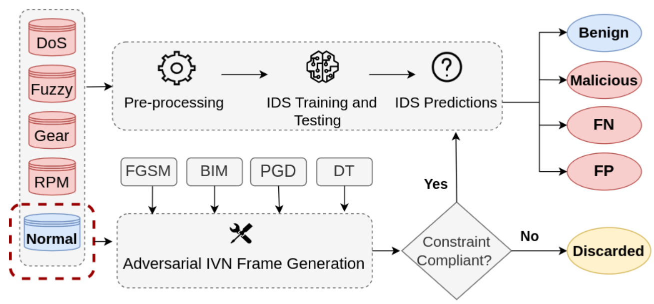

Fig. 2. Adversarial IVN Frame Generation Process

along with a constraint-compliant mask, were fed into each of the four adversarial techniques. The adversarial versions of the normal frames were then fed into five classifiers to examine their effect on IDS performance. We generated these samples under two epsilon values for each adversarial technique: epsilon set to 1 for the first iteration and 5 for the second, with 'epsilon_step' fixed at 0.1. We used the Adversarial Robustness Toolbox [38], a Python library for evaluating and defending against adversarial attacks.

IV. Results

In benign settings, all IDS models achieve a perfect 100% F1 score with minimal FPs. However, under adversarial conditions with varying epsilon values, the IDS models show significant vulnerabilities. Table V details the performance of the five IDS models.

All IDS models demonstrated vulnerabilities when tested on manipulated benign samples, with performance degradation under epsilon values of 1 and 5, especially at epsilon 5. The DT model had the highest attack success rate (ASR) at 89%, despite achieving a 100% F1 score in benign settings. ASR [39], defined as the ratio of successful adversarial samples to total attack attempts (14,237,958 in our case), was also significant for the RF and ET models, with ASRs of 66% and 75%, respectively, contrasting their perfect performance in normal conditions. Conversely, the DNN and XGBoost models showed more robust performance compared to others, with ASRs of 39% and 49%, respectively, indicating greater robustness to the proposed adversarial manipulation. Most misclassified benign frames were identified as spoofing and fuzzy attacks due to their similarity to normal behavior.

FGSM and BIM methods were more effective against DNN and XGBoost compared to other models. This is because gradient-based attacks are particularly effective against differentiable models like DNNs, and XGB's training process involves gradient descent. These methods introduced more noticeable perturbations than PGD. Considering the IVN constraints that limit perturbations and the IDSs' ability to detect minor deviations, FGSM and BIM proved effective in these scenarios. Conversely, attacks designed to exploit the specific structures of models, such as DT attacks, were more effective against DT, RF, and ET. Tree-based ensembles like RF and ET, while more robust than single trees, still inherit vulnerabilities from their constituent decision trees.

In general, our findings highlight that IVN-based IDS are vulnerable to manipulations of benign frames. This underscores the need to evaluate IDS models' performance against both benign and malicious manipulation adversarial attacks. The DNN and XGBoost models demonstrated the best robustness, suggesting they are well-suited for deployment in IVNs. Additionally, our results emphasize the importance of using effective adversarial methods, such as FGSM and DT attacks, to test IVN-based IDS robustness. Addressing this is crucial for developing robust defense and response mechanisms, ensuring the continued efficacy of CAVs.

TABLE V
Comparative Analysis of Targeted IDSs Performance under Benign and Adversarial Settings

Evaluation Parameters		Benign Setting		Adversarial Setting				
Target Model	**Sample Size**	**F1 Score**	**FP**	**Attack**	**Epsilon**	**F1 Score**	**FP**	**ASR**
DNN	14,237,958	100%	914	FGSM	1	99%	298,765	2%
				BIM		99%	292,273	2%
				PGD		99%	216,532	1.5%
				DT		85%	3,625,050	25%
				FGSM	5	76%	5,572,665	39%
				BIM		86%	3,544,486	25%
				PGD		96%	1,147,008	8%
				DT		85%	3,644,552	26%
DT	14,237,958	100%	1	FGSM	1	81%	4,583,605	32%
				BIM		82%	4,283,285	30%
				PGD		97%	849,258	6%
				DT		73%	6,130,534	43%
				FGSM	5	48%	9,739,888	68%
				BIM		67%	7,082,623	50%
				PGD		91%	2,356,413	17%
				DT		19%	12,729,409	89%
RF	14,237,958	100%	0	FGSM	1	62%	7,871,625	55%
				BIM		75%	5,639,331	40%
				PGD		99%	10,984	0.07%
				DT		63%	7,737,320	54%
				FGSM	5	51%	9,427,394	66%
				BIM		65%	7,447,221	52%
				PGD		96%	1,172,766	8%
				DT		56%	8,669,148	61%
ET	14,237,958	100%	0	FGSM	1	98%	514,882	4%
				BIM		98%	514,857	4%
				PGD		100%	0	0%
				DT		40%	10,693,386	75%
				FGSM	5	53%	9,120,210	64%
				BIM		78%	5,123,555	36%
				PGD		99%	110,005	0.8%
				DT		40%	10,693,654	75%
XGBoost	14,237,958	100%	0	FGSM	1	70%	6,584,760	46%
				BIM		82%	4,325,238	30%
				PGD		99%	1,466	0.01%
				DT		99%	124,940	0.9%
				FGSM	5	67%	7,007,221	49%
				BIM		76%	5,578,676	39%
				PGD		98%	400,906	3%
				DT		99%	127,098	0.9%

V. Conclusion

As CAVs become more prevalent, securing their IVNs is crucial. This paper explores vulnerabilities in ML/DL-based IDSs deployed in IVNs to adversarial attacks. Specifically, we shift focus from manipulating malicious frames to appear benign, bypassing detection, to investigating the IDS's susceptibility to benign frames appearing malicious, potentially triggering false alarms. False alarms in safety-critical applications like CAVs can lead to inappropriate responses, risking life-threatening situations, financial damage, and legal liabilities for manufacturers. Our investigation reveals current IDSs are susceptible to adversarial samples (manipulated benign IVN frames) that can trigger false alarms and compromise IDS integrity. Testing five ML/DL-based IDSs with four adversarial techniques resulted in an attack success rate of up to 89%. These findings highlight IDS vulnerabilities not only to manipulated malicious frames but also to benign frames. In the future, robust defense and response mechanisms are urgently needed to enhance IVN security and uphold CAV safety against evolving adversarial attacks.

REFERENCES

[1] T. Limbasiya, K. Z. Teng, S. Chattopadhyay, and J. Zhou, "A systematic survey of attack detection and prevention in connected and autonomous vehicles," *Vehicular Communications*, vol. 37, p. 100515, 2022.

[2] X. Sun, F. R. Yu, and P. Zhang, "A survey on cyber-security of connected and autonomous vehicles (cavs)," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6240–6259, 2021.

[3] R. Bosch GmbH, "CAN Specification," 1991, [online]. Available: http://esd.cs.ucr.edu/webres/can20.pdf.

[4] A. Qayyum, M. Usama, J. Qadir, and A. Al-Fuqaha, "Securing Connected & Autonomous Vehicles: Challenges Posed by Adversarial Machine Learning and the Way Forward," *IEEE Communications Surveys & Tutorials*, vol. 22, pp. 998–1026, 2020.

[5] S.-F. Lokman, A. T. Othman, and M.-H. Abu-Bakar, "Intrusion Detection System for Automotive Controller Area Network (CAN) Bus system: A review," *EURASIP Journal on Wireless Communications and Networking*, p. 184, 2019.

[6] J. Liu, S. Zhang, W. Sun, and Y. Shi, "In-Vehicle Network Attacks and Countermeasures: Challenges and Future Directions," *IEEE Network*, vol. 31, pp. 50–58, 2017.

[7] C. Miller and C. Valasek, "Remote Exploitation of an Unaltered Passenger Vehicle," *Black Hat USA*, 2015.

[8] S. Nie, L. Liu, and Y. Du, "Free-Fall: Hacking Tesla From Wireless to CAN Bus," *Keen Security Lab in Black Hat USA*, 2017, [online]. Available: https://www.blackhat.com/docs/us-17/thursday/us-17-Nie-Free-Fall-Hacking-Tesla-From-Wireless-To-CAN-Bus-wp.pdf.

[9] Z. Cai, A. Wang, W. Zhang, M. Gruffke, and H. Schweppe, "Roadways to Exploit and Secure Connected BMW Cars," *Keen Security Lab in Black Hat USA*, 2019, [online]. Available: https://i.blackhat.com/USA-19/Thursday/us-19-Cai-0-Days-And-Mitigations-Roadways-To-Exploit-And-Secure-Connected-BMW-Cars-wp.pdf.

[10] S. Rajapaksha, H. Kalutarage, M. O. Al-Kadri, A. Petrovski, G. Madzudzo, and M. Cheah, "Ai-based intrusion detection systems for in-vehicle networks: A survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–40, 2023.

[11] B. Lampe and W. Meng, "A survey of deep learning-based intrusion detection in automotive applications," *Expert Systems with Applications*, vol. 221, p. 119771, 2023.

[12] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial Machine Learning," in *4th ACM workshop on Security and artificial intelligence*, 2011.

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing Properties of Neural Networks," *arXiv*, 2014.

[14] P. Papadopoulos, O. Thornewill von Essen, N. Pitropakis, C. Chrysoulas, A. Mylonas, and W. J. Buchanan, "Launching adversarial attacks against network intrusion detection systems for iot," *Journal of Cybersecurity and Privacy*, vol. 1, no. 2, pp. 252–273, 2021.

[15] S. Guo, J. Zhao, X. Li, J. Duan, D. Mu, and X. Jing, "A black-box attack method against machine-learning-based anomaly network flow detection models," *Security and Communication Networks*, vol. 2021, pp. 1–13, 2021.

[16] Y. Pacheco and W. Sun, "Adversarial machine learning: A comparative study on contemporary intrusion detection datasets." in *ICISSP*, 2021, pp. 160–171.

[17] P. Owezarski, "Investigating adversarial attacks against random forest-based network attack detection systems," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2023, pp. 1–6.

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

[19] D. Han, Z. Wang, Y. Zhong, W. Chen, J. Yang, S. Lu, X. Shi, and X. Yin, "Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2632–2647, 2021.

[20] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: an ensemble of autoencoders for online network intrusion detection," *arXiv preprint arXiv:1802.09089*, 2018.

[21] A. Piplai, S. S. L. Chukkapalli, and A. Joshi, "Nattack! adversarial attacks to bypass a gan based classifier trained to detect network intrusion," in *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2020, pp. 49–54.

[22] D. Shu, N. O. Leslie, C. A. Kamhoua, and C. S. Tucker, "Generative adversarial attacks against intrusion detection systems using active learning," in *Proceedings of the 2nd ACM workshop on wireless security and machine learning*, 2020, pp. 1–6.

[23] Z. Lin, Y. Shi, and Z. Xue, "Idsgan: Generative adversarial networks for attack generation against intrusion detection," in *Pacific-asia conference on knowledge discovery and data mining*. Springer, 2022, pp. 79–91.

[24] M. Usama, M. Asim, S. Latif, J. Qadir *et al.*, "Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems," in *2019 15th international wireless communications & mobile computing conference (IWCMC)*. IEEE, 2019, pp. 78–83.

[25] P. T. Duy, N. H. Khoa, H. Do Hoang, V.-H. Pham *et al.*, "Investigating on the robustness of flow-based intrusion detection system against adversarial samples using generative adversarial networks," *Journal of Information Security and Applications*, vol. 74, p. 103472, 2023.

[26] F. Aloraini, A. Javed, O. Rana, and P. Burnap, "Adversarial machine learning in iot from an insider point of view," *Journal of Information Security and Applications*, vol. 70, p. 103341, 2022.

[27] Z. Xiong, H. Xu, W. Li, and Z. Cai, "Multi-source adversarial sample attack on autonomous vehicles," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 3, pp. 2822–2835, 2021.

[28] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks against Machine Learning," 2017.

[29] M. Almehdhar, A. Albaseer, M. A. Khan, M. Abdallah, H. Menouar, S. Al-Kuwari, and A. Al-Fuqaha, "Deep learning in the fast lane: A survey on advanced intrusion detection systems for intelligent vehicle networks," *IEEE Open Journal of Vehicular Technology*, 2024.

[30] L. Yang, A. Moubayed, and A. Shami, "MTH-IDS: A Multitiered Hybrid Intrusion Detection System for Internet of Vehicles," *IEEE Internet of Things Journal*, vol. 9, pp. 616–632, 2022.

[31] H. M. Song, J. Woo, and H. K. Kim, "In-vehicle network Intrusion Detection Using Deep Convolutional Neural Network," *Vehicular Communications*, vol. 21, p. 100198, 2020.

[32] M. Mbow, K. Sakurai, and H. Koide, "Advances in Adversarial Attacks and Defenses in Intrusion Detection System: A Survey," in *Science of Cyber Security (SciSec) 2022 Workshops*, C. Su and K. Sakurai, Eds. Singapore: Springer Nature, 2022.

[33] F. Aloraini, A. Javed, and O. Rana, "Adversarial attacks on intrusion detection systems in in-vehicle networks of connected and autonomous vehicles," *Sensors*, vol. 24, no. 12, p. 3848, 2024.

[34] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv*, 2015.

[35] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial Examples in the Physical World," in *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, 2018.

[36] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," 2019.

[37] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples," 2016.

[38] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. M. Molloy, and B. Edwards, "Adversarial Robustness Toolbox v1.0.0," 2019.

[39] C. Zhang, X. Costa-Pérez, and P. Patras, "Tiki-taka: Attacking and defending deep learning-based intrusion detection systems," in *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2020, pp. 27–39.