ADSecData Platform: An Open-Source Data Platform for Autonomous Driving Cybersecurity

Andrew Roberts[1], Mohsen Malayjerdi[1], Mauro Bellone[1], Raivo Sell[1], Olaf Maennel[3], Mohammad Hamad[2],
Sebastian Steinhorst[2]
[1]Tallinn University of Technology, Estonia
[2]Technical University of Munich, Germany
[3]University of Adelaide, Australia

***Abstract*—Autonomous driving (AD) software needs to be secure, and its decision control must be robust against cyber threats. The development of cybersecurity solutions for legacy and connected vehicles has been supported by an array of open-source datasets, mainly focused on the CAN Bus protocol. There exists a lack of open-source cybersecurity data and community-driven platforms that enable fair and reproducible evaluations of AD algorithms from a cybersecurity perspective and defensive mechanisms. This study addresses this problem by conducting an in-depth analysis of the data ecosystem for AD cybersecurity and introducing an initial open-source data platform, ADSecData. ADSecData offers the community a comprehensive 4-stage method for the creation of AD cybersecurity datasets, along with an initial common dataset. We evaluate the utility of ADSecData through a case study featuring diverse malicious injection attacks, including GPS spoofing, LiDAR point-cloud manipulation, and sensor interference. The results demonstrate the viability of ADSecData in generating AD cybersecurity datasets and supporting community research and development.**

***Index Terms*—Security, Autonomous Driving**

I. Introduction

AD software must be secure, with decision control optimized to ensure robustness against cyberattacks. A key challenge in achieving this goal is the lack of open-source data specifically for AD cybersecurity. Without available data, software designers do not have an immediate understanding of the considerations for secure design required to ensure robustness against cyber threats. In contrast, there are many open-source datasets for safety validation, algorithm optimization, and sensor configuration. Popular examples include KITTI [1], Waymo [2], Baidu Apolloscape [3], Argoverse [4] and NuScenes [5]. Common datasets for safety validation have enabled platforms such as CARLA Leaderboard [6] to establish challenges to benchmark solutions for perception and trajectory planning algorithms. The problem motivation that this research confronts is that AD cybersecurity doesn't have a readily available source of open datasets available to advance research and there is a lack of guidance on how to conduct cybersecurity research to generate datasets for benchmarking.

To confront this problem, we present *ADSecData Platform*, a consolidated platform that provides open-source AD data for cybersecurity. As shown in Fig. 1, ADSecData Platform consists of a data generation process, which is the method used to generate datasets from simulation and real-world experiments. We validate the platform in a case study using the data generation method to create datasets based on an operational autonomous vehicle (AV) program. We demonstrate the utility of our open-source platform to the community in advancing cybersecurity testing to measure and improve the robustness of autonomous driving systems to cyberattacks.

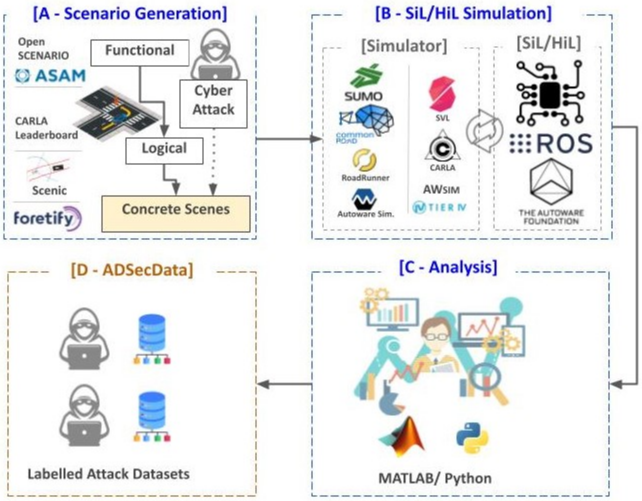

Fig. 1: ADSecData Platform - Data Generation Process.

To construct an AD cybersecurity open-source data platform, we used these guiding questions to establish an understanding of the relationship of AD data to cybersecurity:

1) What data types generated by the AD system are utilized for cyber attack test cases?
2) What is the utility of each data type in enhancing the cybersecurity of AD?
3) What metrics are available to benchmark AD algorithms from a cybersecurity perspective and defense mechanisms?

The major contributions of this research are the follows:

- We present an initial prototype of the *ADSecData Platform*, an open-source platform designed to support AD cybersecurity data.
- We offer detailed guidance on structuring cybersecurity testing frameworks to facilitate the generation of datasets for the research community.
- We contribute to the AD cybersecurity community by providing an initial common dataset and defining key challenges to focus future research and development efforts.

II. Autonomous Vehicle Cybersecurity Data

The emerging field of automotive cybersecurity research over the last decade has focused predominantly on the CAN Bus protocol, connected vehicle protocols, electrical and embedded hardware (such as wireless controllers and Bluetooth), and in-vehicle software systems (e.g., infotainment systems). To support the development of defensive technologies and the secure design of communication protocols and software, numerous open-source datasets of automotive telemetry have been created. These datasets primarily address legacy and connected vehicle technologies, with a strong emphasis on the CAN Bus protocol. However, there is a significant lack of open-source cyberattack datasets specific to AD technology. Developing such datasets and promoting the exchange of open-source data are critical steps toward advancing the still-maturing field of AD cybersecurity.

A. Autonomous Vehicle Data

AD systems generate a vast amount of data from diverse hardware and system components. We classify AD data into four major sub-categories of data sources: *sensing*, *system*, *network*, and *vehicle dynamics*. For each data source, we discuss its value for software development, cybersecurity, and its availability.

1) Sensing: Sensing data is produced by advanced sensors in the AD system, including LiDAR, cameras, ultrasonic radar, and global navigation systems (GPS, GLONASS, Baidu, Galileo). This data is critical for mapping the driving environment, perception, and localization. However, one of the key challenges with sensor data is the high data rate generated by autonomous vehicles. Xu et al. [7] estimated that diverse sensors could generate approximately four terabytes of data per day. The transmission of LiDAR and high-definition camera frames from onboard sensors to edge data logging servers further complicates data collection. Although compression techniques are available to optimize transmission efficiency, there is limited understanding of how these methods impact cybersecurity research in computer vision and perception.

Software Development Value: Sensing data is used by AD software designers to train and optimize algorithms for SLAM, object detection and tracking, sensor fusion, and semantic segmentation. One of the many examples of progress in this area is the CARLA Autonomous Driving Leaderboard [6], which is a platform used for the development of AD agents.

Cybersecurity Value: Sensing data can be used to assess vulnerabilities of AD software to adversarial examples and generate new attack models for adversarial examples. Select examples include:

- LiDAR point cloud manipulation [8]
- Adversarial examples for camera perception neural networks. [9]
- Light manipulation attacks on camera hardware and driving objects (road signs, etc.) [10]
- Fuzzing and parameter manipulation attacks against AD algorithms (Object Detection, Sensor Fusion) [11]
- GPS Spoofing causes uncertainties in trajectory planning algorithms. [12]

Defensive technologies can also be developed from sensing data, these include:

- Kalman filters and ML detection solutions to filter noise from data manipulation attacks. [13]
- Physical intrusion detection solutions which fingerprint patterns of noise from adversarial activity. [14]
- Improvements to the security of ML models to protect against ML evasion and training data poisoning attacks.

Data Availability: Open-source cybersecurity datasets for sensing, of which there are very few, predominantly focus on camera-based perception and neural networks for perception algorithms. Available datasets include:

- *Natural Denoising Diffusion Attack (NDDA) dataset* [15]
- *SlowTrack: Camera-based perception latency attack dataset* [16]

2) System: System data consists of data from the onboard software systems of the AD system. These include the firmware, operating system, application software, and real-time operating systems used in the electronic/embedded components such as the electronic control units (ECUs) and microelectronic control units (MCUs).

Software Development Value: System data is used by software developers to debug errors and understand application performance and functionality. Crucial for AV developers is to understand the performance and reliability of the AD software (Autoware, Nvidia Drive, Apollo) and middleware (Robotic Operating System (ROS), Cyber RT).

Cybersecurity Value: System data is used for vulnerability and exploit analysis. Activities that are included in this description include reverse engineering firmware, code analysis, taint-analysis, and fuzz testing.

Data Availability: System datasets are generally available from the manufacturer. These are then used for vulnerability and exploit analysis. Cybersecurity datasets are rare as the responsible disclosure process usually results in the removal and updating of new software. Examples of a cybersecurity system artifact are the following:

- *Kia OFFensivE Exploit (KOFFE) Metasploit module* [17]
- Mazda Infotainment USB attack [18]

3) Network: Network data consists of data produced from the AV internal and external networks. CAN Bus is the network that predominates in in-vehicle communication between ECUs and handles critical real-time functions such as braking and steering actuation. Automotive Ethernet is gaining in popularity and is mostly used for drive-by-wire communication. Other communication, such as MOST, is used for infotainment systems, and LIN can be found in more upmarket vehicle classes. The difficulty in providing CAN (and most other in-vehicle protocols) datasets is that CAN is used in a proprietary format by vehicle manufacturers. To decipher the meaning of CAN messages, either the manufacturer's diagnostic tool is required or knowledge to reverse engineer CAN messages from the investigation of firmware and system manuals.

For legacy and connected vehicles, great progress has been made, and there exist many available datasets and tools to help with the CAN message extraction process [19]. However, to our knowledge, no CAN cybersecurity-specific datasets exist for AD technology. Reasons for this could be the enhanced commercial sensitivity of AD technology, a more diverse range of AV manufacturers, the implementation of encrypted messaging with CAN-FD, and the cutting-edge nature of AD technology. Other network concepts typical in AD architectures include Vehicle-to-vehicle (v2v) and vehicle-to-everything (v2x), which use wireless and cellular connectivity for connectivity. Different application layer protocols are used for distinct purposes, these may include MQTT for vehicle on-board unit (OBU) to edge communication and Cooperative v2x (C-V2x) protocols that including basic safety messages (BSM) for cooperative perception and intelligent feedback for decision-making.

Cybersecurity research in this field is well-developed, and many available studies investigate attack models to the integrity of cooperative vehicular messages and the availability of networks that support vehicle data processing and cooperative communication.

Software Development Value: For software developers, network datasets can assist in understanding system interconnection and latency of data flow through situational awareness data to control actions decided by AD software and physical processes made by actuation.

Cybersecurity Value: Network datasets are primarily used for defensive intrusion detection solutions. Network datasets also aid in developing new attack strategies (DDoS, Replay, etc.) and fuzzing strategies to test the robustness of communication architectures. Lately, as more CAN cybersecurity datasets are available, research has focussed on ML and AI solutions for automated attack detection and fuzzing [20]. Within AD architectures, network data is utilized to evaluate the security aspects of cooperative driving, such as message trust and authentication. Perhaps the greatest contribution of cybersecurity CAN datasets has been the increase in attention brought by attacks, which demonstrate the feasibility of cyber attacks to manipulate safety-critical functions such as braking, steering, and acceleration. Recognition of these threats has seen the development of security within automotive software architectures (AUTOSAR Adaptive) and new zonal communication architectures for in-vehicle network communications.

Data Availability: Open-Source CAN hacking datasets exist for legacy and connected vehicles, a sample of this long list include:

- *Car-Hacking-Dataset* [21] [22]
- *Survival Analysis Dataset* [23]
- *CAN-Train-And-Test Dataset* [24] [25]
- CANet Dataset [26]
- CrySyS Dataset [27]
- *CIC IoV 2024 Dataset* [28]
- *CAN-MIRGU Dataset** [29]

**The CAN-MIRGU dataset is generated from a vehicle with AD capabilities, however, these capabilities are not detailed due to privacy reasons and the AD functions are deactivated for safety reasons.* For V2X and V2V selected datasets include:

- *Simulated VANET Attack Dataset* [30]
- *Simulated VANET Attack Dataset* [31]

4) Vehicle Dynamics: Vehicle dynamics data include body physical movement (lateral and longitudinal pose, yaw, etc.), acceleration, braking, and steering actuation. Vehicle dynamics is crucial for a software developer and cybersecurity engineer to understand how behavior at a system level affects the vehicle. Existing cyber attack research, which focuses on vehicle dynamics, predominantly concerns itself with providing artifacts such as docker images of the attack simulation and the code base for adversarial examples and fuzzing tools. A limitation of this approach is that it requires a custom configuration of the attack in the user environment and an understanding of the vehicle model and metrics engine for data output used in the original research.

Software Development Value: This data is crucial for control algorithm designers to assess the robustness of control and trajectory planning algorithms. Software developers and control designers will use vehicle dynamics data for backstepping and back-propagation of the AD control software.

Cybersecurity Value: Vehicle dynamics data enables a greater understanding of the effect of cyber attacks on vehicle behavior. The utility of vehicle dynamics data includes research and development of physical intrusion detection system solutions and root cause analysis.

Data Availability: We are not aware of any datasets for vehicle dynamics in the context of cybersecurity.

B. Gaps in Autonomous Vehicle Datasets

Our exploration of diverse AD data types and their usage in cybersecurity has identified a number of limitations:

- **Lack of a consolidated research data platform.** Datasets are distributed across GitHub accounts and research papers. There is a lack of consolidation of datasets that would enable security research across the AD technology stack.
- **Siloed research.** Defensive mechanisms are often developed based on a single data type (e.g., CAN, Camera, etc.). The lack of availability of other data sources and an understanding of how this data impacts vehicle dynamics and propagates through the AD system results in the

creation of defense mechanisms that lack system-level validation.

- **Lack of cybersecurity data:** There is a lack of data for cybersecurity, and in some of the sub-categories explored, there is, to our knowledge, no data available. The available datasets overwhelmingly consist of legacy and connected vehicles.

III. ADSecData Platform

In developing a method for generating cybersecurity data for AD systems, the significant change from legacy vehicles is the focus on vehicle behavior. As the vehicle is controlled by software and algorithms, it is important to understand the effect of cyber activity on the vehicle and its implications for decision control. In addition to attacks that directly target AD technologies, such as advanced sensors, attacks on network and system components can have a downstream effect on autonomous control. The ADSecData Platform (shown in Fig. 1) follows a four-stage process for generating data.

A. Scenario Generation

Scenario-based testing (SBT) involves evaluating the performance of a module or the full AD pipeline (perception, localization, planning, and decision-control) to perform its task during a specified driving scenario. Since the performance of algorithms can vary under diverse scenarios, SBT has become the standardized approach for AD algorithm safety validation and verification testing [32]. Cybersecurity represents an edge and corner case for SBT. For the ADSecData methodology, we propose that scenario generation is a crucial step for cybersecurity, as it is essential to understand whether the effect of a cyber attack on the vehicle differs based on the scenario. Since scenario libraries for AD cybersecurity testing are not available, our methodology recommends using safety validation testing libraries (such as ASAM OpenScenario, etc.) and customizing the scenarios with attack models.

B. Simulation/Test Environment

As the task of driving can encounter many diverse scenarios, simulation is the only feasible mechanism to incorporate large-scale testing agilely. Cybersecurity testing should be aligned with safety validation testing, where the choice of test environment is based on evaluating the algorithm's ability to perform tasks. This is part of a testing process that uses regression testing to map scenario test sets from simulation test environments to real-world proving grounds. Within the ADSecData platform, we recommend using low-fidelity test environments for large-scale testing of driving logic, high-fidelity test environments to include testing of advanced sensors (such as LiDAR, Camera, etc.), and real-world proving grounds. Another factor influencing the integrity of cybersecurity data is the tendency of automotive cybersecurity practitioners to provide singular datasets based on attack type. Due to the experimental nature of AD algorithms, sufficient tests need to be run to ensure that anomalous vehicle behavior is caused by cyber activity and not system errors or a lack of algorithm optimization.

Another key aspect of the simulation/test environment stage is defining metrics and configuring the format of output data. Safety metrics and vehicle dynamic parameters are applied to quantify the impact of cyber activity on the vehicle. Cybersecurity labels include details such as the attack initiation during the scenario, attack parameters (e.g., sensor interference noise level, GPS positioning offset), and their corresponding weighting.

TABLE I: Requirements for ADSecData

Category	Requirement
Documentation	• Dataset should be accompanied by general documentation describing content and origin. • Documentation should include description of the attacks in the dataset and how they were executed/recorded. • Documentation should include description of the features (e.g., origin, meaning, range) and their physical context (e.g., how vehicle speed, engine speed and gear are related).
Labels	• Each entry in the dataset may be given a label for identifying whether that entry is benign or an attack.
Parseability, correctness and consistency	• Data should be stored in an appropriate machine/humanreadable format (e.g., PCAP or CSV rather than SQL databases) • All entries should be correctly formatted (e.g., no corrupt entries) • use a single data format for all entries
Age, Size, Objective	• Dataset should not be legacy (> 5 years old etc.) and consist of a balance between benign and cyber attack data.
Completeness	• Dataset should be complete in the sense that no key features or entries have been discarded.
Transformation and anonymization	• Data should not be irreversibly transformed (changing timestamps etc.) and not be anonymised to the point that it bias' detection mechanisms.
Dataset and Attack Realism	• Dataset should include diverse attacks and not be wholly based on synthetic data.

C. Analysis

The analysis stage involves interrogating the data to assess its integrity and accuracy, ensuring consistency with the experimentation performed. Popular tools, including MATLAB and Python, are used to plot data, visualize patterns, and analyze trends. For example, analyzing a dataset from the trajectory planning module could generate trajectory maps to visualize the vehicle's path and highlight any deviations from the reference path. Analysis is a crucial activity for identifying problems with the experimentation process and evaluating the data quality.

D. ADSecData

Data should be benchmarked for measurement and comparison. The benchmarks for automotive cybersecurity datasets from Vahidi et al. [33] systematic evaluation of automotive intrusion datasets serve as a good starting point. We utilize their data requirements to develop the ADSecData Platform and data readiness labels. Table. I provide the requirements for ADSecData datasets, the classification is provided in more detail on the platform website.

IV. ADSecData Case Study

A. Target Autonomous Vehicle

The target vehicle is an AV for public transportation, that is an autonomous electric vehicle (AEV). The shuttle operates at Level 4 autonomy (high automation), meaning that it can handle most driving tasks without human intervention in predefined areas, and it is equipped with advanced LiDAR, radar, cameras, and GPS systems to navigate safely and carry out perception tasks in an urban environment. Its software

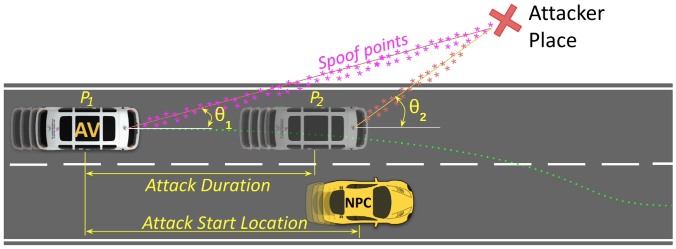

Fig. 2: Attack Case 1 Threat Model.

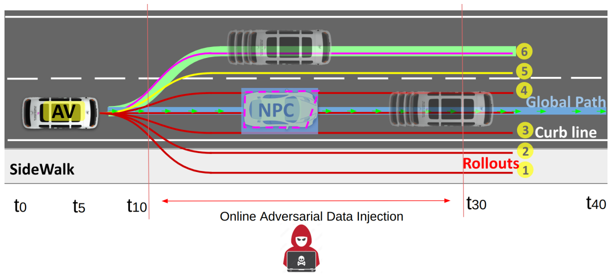

Fig. 3: Attack Case 2 & 3 Threat Model.

backbone is based on ROS and Autoware, controlling all the driving functionalities and implementing the driving dynamic model of the vehicle.

B. Scenarios

Our initial dataset consists of 4 attack cases conducted during diverse driving scenarios.

Attack Case 1 - LiDAR point-cloud manipulation: The LiDAR point-cloud manipulation attack, as shown in Fig. 2, consists of an adversary with a LiDAR capable of injecting malicious LiDAR point clouds into the LiDARs of the AV. This attack is conducted whilst the AV is attempting an overtaking maneuver.

Attack Case 2 - Position Offset: Attack Case 3 - Message Delay: The attacker creates a spoofed ROS topic which is able to deliver malicious input data of the `Current_Pose` (longitude, latitude, and velocity) to all the nodes of the local planning module. The data manipulation is injected online/dynamically during the critical overtaking manoeuvre involving the AV and NPC (Non-playable character). Figure 3 displays the critical driving scenario and the time frames in which the manipulated `Current_Pose` data is injected into the local planning pipeline cost estimation. The red dashed lines in Fig. 3 represent the roll-outs, and the green highlighted, denoting the selected motion-path.

For the manipulation of the `Current_Pose` data, we introduce a deviation to lateral and longitudinal pose. For the lateral pose data, the sensitivity deviation introduced was structured as follows:

- Attack Case 2a: 0.16%
- Attack Case 2b: 0.33%
- Attack Case 2c: 0.5%

This range represents a slight perturbation of pose to a 1m deviation. The longitudinal pose data sensitivity deviation range was structured as follows:

- Attack Case 2d: 0.33%
- Attack Case 2e: 0.66%
- Attack Case 2f: 1.00%

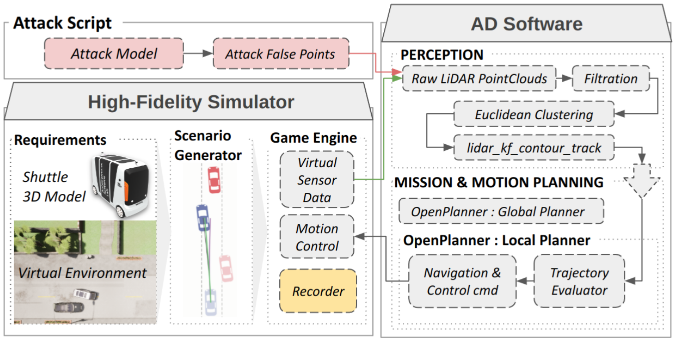

Fig. 4: Architecture of the testing platform.

This range is the same as the longitudinal deviation. The difference in percentage comes from the difference in coordinate values of lateral and longitude. The lateral value is almost double that of the longitudinal, and therefore, the percentage is doubled.

This attack scenario involves introducing a time-delay into the messages of the Current_Pose topic communicating to the nodes of the local planning module.

We introduced a message delay when the AV passes 2m in front of the vehicle that it is passing in the lateral direction. We introduce 3 different time delays in the message:

- Attack Case 3a: 0.3 seconds
- Attack Case 3b: 0.6 seconds
- Attack Case 3c: 1.0 seconds

The message frequency is approximately 50hz, so this is a message every 20 milliseconds. We chose the above range of deviation of time-delay as it enabled a spectrum of a message from the delay from approximately 15, to 50 messages.

Attack Case 4 - GPS Spoofing: The attack model of GPS spoofing involves an adversary using a transmitter near the AV and interferes with the GPS signals being transmitted.

C. Simulation/Test Environment

Attack Case 1 was conducted in the high-fidelity CARLA simulator [34]. In this study, we use Carla 0.9.13 as the high-fidelity simulator. Figure 4 illustrates the requirements for the high-fidelity simulator to conduct simulation testing, which are two components, the digital twin of the target AV and the virtual replication of our target environment. These replicated components help us to gain more accurate results of the proposed platform [35]. The AV digital twin is a 3D model of the target real-world world AV shuttle, designed in Blender, a graphical 3d modeling software, and imported and built in Unreal for deployment in CARLA. This model uses the same dimension and sensor configuration (model, position, and orientation) from the real AV shuttle. The environment digital twin, in our case, is identical to the location where the vehicle operates.

This simulation setup was implemented on a desktop computer with the following configuration:

- Intel® Core™ i7-11700K @ 3.60GHz × 16 cores
- NVIDIA GeForce RTX 3080 10 GB
- RAM: 128 GB

Attack Case 2 and 3 were conducted in a low-fidelity simulator. To accelerate the testing, we bypassed the sensing and detection nodes of the algorithm and focused on the planning part by utilizing the low-fidelity simulation feature provided by Autoware.ai and Openplanner. The low-fidelity simulation uses the open-planner 2.5 control algorithm. It provides simulated localization and detection data for the planning nodes and receives the actuation commands to simulate the AV kinematics. This process runs faster due to the low-detail environment required for the simulation and the lack of the process to simulate the sensors.

Attack Case 4 dataset was generated from the real-world vehicle. GPS spoofing activity occurred during a point-in-time of a 3 month trial of AVs in a city in Northern Europe.

D. Analysis

The data output parameters were defined based on safety, vehicle dynamics and security criteria. A sample of these includes safety criteria, mission success, violation, break status, and distance-to-collision. Vehicle dynamics included steer, yaw, lateral and longitudinal position. Security criteria include 2 labels, *is_attack* denoting when the attack is occurring and *cyber_weight*, which denotes the level of sensor noise manipulation.

E. ADSecData

The 4 attack case scenarios datasets were generated as a .csv files. Each attack includes a corresponding benign (no attack) dataset to benchmark the stability of the AD algorithms under the given driving scenario. *Attack Case 1* included over 1200 simulations. *Attack Case 2 and 3* included over 900 simulations collectively. The data is available at this link: ADSecData Platform (*To note: this website is abridged, anonymised version*)

V. Discussion

The case study provides a starting point for the development of a common dataset for the community to perform fair and reproducible evaluations of AD algorithms for cybersecurity and defensive mechanisms. The datasets generated from the 4 attack cases demonstrate the importance of following the 4 stage ADSecData method where particular careful consideration is taken in the definition of data output parameters and experimental evaluation analysis. For the development of the ADSecData platform, community challenges, and a roadmap are fundamental.

A. Community Challenges

These are the first tranche of community challenges that we recommend for the ADSecData platform:

Ch1 Performance and Accuracy of Semantic Fuzzing Tools.
Ch2 Intrusion Detection of Semantic AD Sensor Attacks.
Ch3 Robust Sensor Fusion Algorithms.
Ch4 Robust and Resilient Trajectory Planning Algorithm.

We see these challenges as of most immediate importance and value for the community. Furthermore, we would like to see the community use the ADSecData platform to generate a seed corpus for guided semantic data fuzzing tools. As large language models (LLMs) are gaining in popularity, another foreseeable use would be to apply LLMs to ADSecData to generate scenarios for cybersecurity testing. As AD cybersecurity lacks a common scenario library, the generation of cybersecurity scenarios would help to close this gap. Finally, IDS solutions for attacks on the AD sensors are essential to mitigate the risk to AD control. There needs to be more data to understand the profile of cyber attacks compared to emergency and safety actions from edge and corner cases.

B. Future Roadmap

The short-term aims of the ADSecData platform are to add more datasets from all 4 sub-categories of data types and different vehicle classes and increase the community's awareness of the platform. There will be a need to improve the development of both the front-end and back-end platforms to enable secure data sharing and a more intuitive user experience. Longer-term aims include a need to investigate metrics for intrusion detection solutions for AD, which is an AI-based system. Traditionally, MITRE ATT&CK is used for benchmarking IDS solutions, and MITRE has a framework for AI, MITRE ATLAS. It would be interesting to evaluate how this would work in a practical use-case for AD.

VI. Related Work

There have been attempts by the community to build common infrastructure for AV cybersecurity testing. PASS [36] and Simutack [37] are community simulation testing platforms. Whilst these platforms are valuable to the community and enable accessibility of simulation testing to researchers, the usage of community simulation testing platforms is limited as real-world operators tend to use their own customised platforms. Furthermore, neither of these studies focused on the data aspect of cybersecurity testing as part of their scope. Lauinger et al. [38] developed an attack data generation framework for AVs. Our work enhances this contribution by integrating the concepts of scenario generation and simulation and testing environments for data generation.

From a community data sharing perspective, there are initiatives such as Platform for Innovative use of Vehicle Open Telematics (PIVOT) [39], which is a U.S National Science Foundation project to create a open-source portal for vehicle telemetry data in the context of cybersecurity. However, as of writing this portal was unavailable.

As aforementioned in Section. II, there exists a diversity of datasets for legacy and connected vehicles. There are also the studies of Vahidi et al. [33], Lampe & Meng [25] and Lee et al. [40] which evaluate cybersecurity data of legacy and connected vehicles for intrusion detection. However, to our knowledge, there are no existing contributions that focus on the autonomous technology stack of AVs.

VII. Conclusion

In this work, we present an open-source data platform for AD cybersecurity, the ADSecData platform. Further, we detail a 4-stage method for the generation of AD cybersecurity datasets. We demonstrate the utility of the ADSecData platform through the generation of open-source datasets for four diverse AD cybersecurity attacks, which we provide to the community. The ADSecData platform is available to the community and will continue to develop according to the challenges and roadmap presented in this study.

Acknowledgment

This work is partly supported by the Horizon Europe projects CyberSecDome (Agreement No.: 101120779) & XTRUST-6G (Agreement No.: 101192749) and the European Union and Estonian Research Council via project TEM-TA5.

References

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] P. Wang, X. Huang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloscape open dataset for autonomous driving and its application," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[4] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.

[5] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[6] CARLA, "Carla autonomous driving leaderboard," 2024. [Online]. Available: https://leaderboard.carla.org/

[7] W. Xu, H. Zhou, N. Cheng, F. Lyu, W. Shi, J. Chen, and X. Shen, "Internet of vehicles in big data era," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 1, pp. 19–35, 2018.

[8] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, "Adversarial sensor attack on lidar-based perception in autonomous driving," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '19. Association for Computing Machinery, 2019.

[9] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[10] T. Sato, S. H. Bhupathiraju, M. Clifford, T. Sugawara, Q. A. Chen, and S. Rampazzi, "Wip: Infrared laser reflection attack against traffic sign recognition systems," *ISOC Symposium on Vehicle Security and Privacy (VehicleSec)*, 2023. [Online]. Available: https://par.nsf.gov/biblio/10427118

[11] R. S. Hallyburton, Y. Liu, Y. Cao, Z. M. Mao, and M. Pajic, "Security analysis of Camera-LiDAR fusion against Black-Box attacks on autonomous vehicles," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 1903–1920. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/hallyburton

[12] R. Jiao, J. Bai, X. Liu, T. Sato, X. Yuan, Q. A. Chen, and Q. Zhu, "Learning representation for anomaly detection of vehicle trajectories," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 9699–9706.

[13] Y. Jia, Y. Lu, J. Shen, Q. A. Chen, H. Chen, Z. Zhong, and T. Wei, "Fooling detection alone is not enough: Adversarial attack against multiple object tracking," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rJl31TNYPr

[14] R. Quinonez, J. Giraldo, L. Salazar, E. Bauman, A. Cardenas, and Z. Lin, "Savior: securing autonomous vehicles with robust physical invariants," in *Proceedings of the 29th USENIX Conference on Security Symposium*, ser. SEC'20. USA: USENIX Association, 2020.

[15] T. Sato, J. Yue, N. Chen, N. Wang, and Q. A. Chen, "Intriguing Properties of Diffusion Models: An Empirical Study of the Natural Attack Capability in Text-to-Image Generative Models," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] C. Ma, N. Wang, Q. A. Chen, and C. Shen, "SlowTrack: Increasing the Latency of Camera-Based Perception in Autonomous Driving Using Adversarial Examples," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4062–4070.

[17] I. M. G. Costantino, M. De Vincenzi, "A vehicle firmware security vulnerability: an ivi exploitation." *J Comput Virol Hack Tech*, vol. 20, pp. 681,696, 2024.

[18] Turla, "Mazda getinfo attack," 2017. [Online]. Available: https://github.com/shipcod3/mazda_getInfo

[19] M. D. Pesé, T. Stacer, C. A. Campos, E. Newberry, D. Chen, and K. G. Shin, "Librecan: Automated can message translator," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2283–2300. [Online]. Available: https://doi.org/10.1145/3319535.3363190

[20] N. Alkhatib, M. Mushtaq, H. Ghauch, and J.-L. Danger, "Can-bert do it? controller area network intrusion detection system based on bert language model," in *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, 2022, pp. 1–8.

[21] E. Seo, H. M. Song, and H. K. Kim, "Gids: Gan based intrusion detection system for in-vehicle network," in *2018 16th Annual Conference on Privacy, Security and Trust (PST)*, Aug 2018, pp. 1–6.

[22] 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.

[23] M. L. Han, B. I. Kwak, and H. K. Kim, "Anomaly intrusion detection method for vehicular networks based on survival analysis," *Vehicular Communications*, vol. 14, pp. 52–63, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214209618301189

[24] B. Lampe and W. Meng, "can-train-and-test: A curated can dataset for automotive intrusion detection," *Computers & Security*, vol. 140, p. 103777, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404824000786

[25] ——, "can-train-and-test: A new can intrusion detection dataset," in *2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall)*, 2023, pp. 1–7.

[26] M. Hanselmann, T. Strauss, K. Dormann, and H. Ulmer, "Canet: An unsupervised intrusion detection system for high dimensional can bus data," *IEEE Access*, vol. 8, pp. 58 194–58 205, 2020.

[27] A. Gazdag, R. Ferenc, and L. Buttyán, "Crysys dataset of can traffic logs containing fabrication and masquerade attacks," *Scientific Data*, vol. 10, 2023.

[28] E. C. P. Neto, H. Taslimasa, S. Dadkhah, S. Iqbal, P. Xiong, T. Rahman, and A. A. Ghorbani, "Ciciov2024: Advancing realistic ids approaches against dos and spoofing attack in iov can bus," *Internet of Things*, vol. 26, p. 101209, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660524001501

[29] S. Rajapaksha, G. Madzudzo, H. Kalutarage, A. Petrovski, and M. O. Al-Kadri, "Can-mirgu: A comprehensive can bus attack dataset from moving vehicles for intrusion detection system evaluation," in *Symposium on Vehicles Security and Privacy. Internet Society*, 2024.

[30] S. Iqbal, P. Ball, M. H. Kamarudin, and A. Bradley, "Simulating malicious attacks on vanets for connected and autonomous vehicle cybersecurity: A machine learning dataset," in *2022 13th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, 2022, pp. 332–337.

[31] F. Gonçalves, B. Ribeiro, O. Gama, J. Santos, A. Costa, B. Dias, M. J. Nicolau, J. Macedo, and A. Santos, "Synthesizing datasets with security threats for vehicular ad-hoc networks," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[32] Y. Huai, Y. Chen, S. Almanee, T. Ngo, X. Liao, Z. Wan, Q. A. Chen, and J. Garcia, "Doppelgänger test generation for revealing bugs in autonomous driving software," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2023, pp. 2591–2603.

[33] A. Vahidi, T. Rosenstatter, and N. I. Mowla, "Systematic evaluation of automotive intrusion detection datasets," in *Proceedings of the 6th ACM Computer Science in Cars Symposium*, ser. CSCS '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3568160.3570226

[34] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[35] M. Malayjerdi, V. Kuts, R. Sell, T. Otto, and B. C. Baykara, "Virtual simulations environment development for autonomous vehicles interaction," in *ASME International Mechanical Engineering Congress and Exposition*. American Society of Mechanical Engineers, 2020.

[36] Z. Hu, J. Shen, S. Guo, X. Zhang, Z. Zhong, Q. A. Chen, and K. Li, "Pass: A system-driven evaluation platform for autonomous driving safety and security," *NDSS Workshop on Automotive and Autonomous Vehicle Security (AutoSec)*, 2025. [Online]. Available: https://par.nsf.gov/biblio/10359464

[37] A. Finkenzeller, A. Mathur, J. Lauinger, M. Hamad, and S. Steinhorst, "Simutack - an attack simulation framework for connected and autonomous vehicles," in *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, 2023, pp. 1–7.

[38] J. Lauinger, A. Finkenzeller, H. Lautebach, M. Hamad, and S. Steinhorst, "Attack data generation framework for autonomous vehicle sensors," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2022, pp. 128–131.

[39] P. Project, "Platform for innovative use of vehicle open telematics," 2024. [Online]. Available: https://pivot-auto.org/

[40] S. Lee, W. Choi, I. Kim, G. Lee, and D. H. Lee, "A comprehensive analysis of datasets for automotive intrusion detection systems," *Computers, Materials and Continua*, vol. 76, no. 3, pp. 3413–3442, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1546221823000516