Towards a Hybrid Generative AI-Based Framework for Automotive Cybersecurity: A Feasibility Study

André Gheventer
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
gheventer@ufrj.br

Claudio Miceli de Farias
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
cmicelifarias@cos.ufrj.br

Rafael Maiani de Mello
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
rafaelmello@ic.ufrj.br

Abstract

Ensuring cybersecurity in Software-Defined Vehicles (SDVs) requires continuous operational observability; however, Vehicle Security Operations Centers (VSOCs) face two critical challenges: the scarcity and imbalance of real-world attack data for training Intrusion Detection Systems, and the cognitive overload on analysts interpreting raw vehicular logs. This paper proposes a Hybrid Generative AI-Based (GenAI) framework to address these issues by integrating a Conditional Tabular Generative Adversarial Network (GAN) for data augmentation and a lightweight Large Language Model (LLM) for alert interpretation. A first feasibility study using a GAN and LLM demonstrates the potential of GenAI but also reveals critical limitations in off-the-shelf models, including semantic integrity violations in generated packets and hallucination risks in alert interpretation, underscoring the need for robust architectural orchestration and tailored software engineering practices in critical automotive environments.

Keywords

Generative AI, Software-Defined Vehicle, GAN, IDS, Security, VSOC

1 Introduction

The global automotive industry is undergoing a critical transformation driven by the rise of Software-Defined Vehicles (SDVs) and the massive adoption of Internet of Things (IoT) technologies. Modern vehicles have transitioned from mechanical systems into intelligent computing nodes, continuously sharing real-time data within the Internet of Vehicles (IoV) ecosystem [2]. While this connectivity brings undeniable benefits, it also exposes vehicles to a wide range of cyberattacks [1]. Recent vulnerabilities identified in vehicles already in operation—ranging from physical Controller Area Network (CAN) bus injections (CVE-2025-6785) to cryptographic flaws that expose sensitive telemetry logs and location data in infotainment systems (CVE-2025-7020) [3]—demonstrate that factory-level pre-production security practices remain insufficient, highlighting the critical need for agility to rapidly detect and mitigate emerging threats on the road. To address this dynamic threat landscape, global regulations such as UN R155 [13] and standards like ISO/SAE 21434 [9] establish strict requirements for continuous cybersecurity monitoring and incidents response capabilities throughout the vehicle's lifecycle, typically operationalized through a Vehicle Security Operation Center (VSOC) [8].

Despite their importance, operationalizing a VSOC faces significant technical barriers. First, threat detection, particularly for unknown attacks[1], is hindered by a fundamental obstacle: the scarcity and imbalance of representative training data for Intrusion Detection Systems (IDS) from real-world cyberattacks [11]. Obtaining such data in controlled physical environments is difficult, and datasets quickly become obsolete as attack vectors evolve [14]. Second, the heterogeneity and low readability of raw vehicular data, such as logs from the Controller Area Network (CAN)—the central nervous system of in-vehicle communication—impose a high cognitive load on human analysts [2]. Identifying anomalies requires correlating hexadecimal payloads, message frequencies, and temporal behaviors, which significantly delays incidents response.

To address these challenges, Generative Artificial Intelligence (GenAI) has emerged as a promising approach. Generative Adversarial Networks (GANs) can synthesize high-fidelity tabular data to augment training datasets [4], while Large Language Models (LLMs) can act as virtual analysts, translating complex logs into highly interpretable natural language diagnostics [10]. However, adopting off-the-shelf GenAI solutions in cyber-physical ecosystems remains complex. Recent empirical observations indicate a lack of industrial readiness and friction in transitioning these technologies to real-world scenarios [1, 7] due to systemic limitations, such as critical hallucinations and a lack of semantic integrity when handling specific vehicular protocols [10].

Building upon the need for system-level orchestration, this paper proposes a hybrid GenAI-based framework to support automotive cybersecurity. To validate the feasibility of this architecture, we conducted a first empirical study that employed a Conditional Tabular GAN (CTGAN) to generate CAN bus traffic, a Random Forest-based IDS for anomaly detection, and a lightweight LLM for alert interpretation. The main contribution of this work is to expose engineering challenges that must be addressed before GenAI can be safely deployed in VSOC environments. To this end, rather than solely presenting positive metrics, our investigation uncovers the structural and semantic limitations of off-the-shelf GenAI models, demonstrating that robust architectural interventions and tailored software engineering practices are required to ensure reliability.

The remainder of this paper is organized as follows. Section 2 details the proposed framework. Section 3 presents the feasibility study, including the methodology, setup, and preliminary results. Section 4 discusses the broader research context, outlines its limitations, and proposes future work.

2 The Hybrid GenAI-Based Framework

The proposed framework is a continuous pipeline for observability and decision support, designed to integrate with simulation environments and real-world operations. As depicted in Figure 1, the framework integrates a preliminary sanitization stage followed by three interdependent phases:

[1] An attack that exploits a previously unknown hardware, firmware, or software vulnerability. https://csrc.nist.gov/glossary/term/zero_day_attack/.

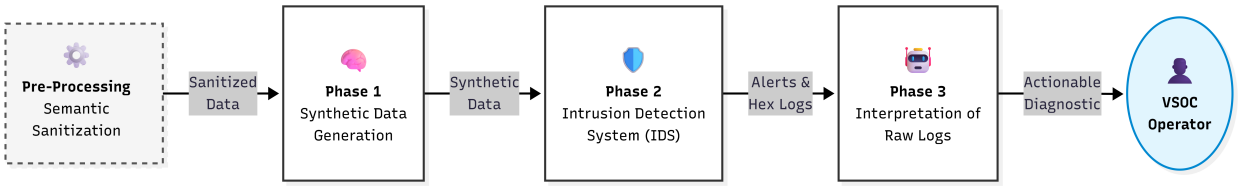

Figure 1: Proposed Hybrid Architecture for VSOCs, illustrating the flow from data augmentation to alert interpretation.

Pre-Processing - Semantic Sanitization. In real-world environments, raw vehicular datasets can exhibit varied formats, anomalies, and inconsistencies, ranging from legitimate variable-length messages to corrupted data streams due to signal noise. Direct extraction without prior analysis can create structural misalignments and spread noise into the learning phase. Therefore, the main goal of pre-processing is to sanitize the data, ensuring structural and protocol integrity before feeding it to the subsequent Phase 1.

Phase 1 - Synthetic Data Generation. Taking the sanitized data from the previous stage as input, this phase uses a CTGAN to augment the dataset by generating synthetic attack scenarios and anomalous CAN bus traffic. This is necessary because an IDS requires large volumes of balanced data to learn complex patterns without bias. The goal is to generate synthetic datasets that preserve statistical fidelity and the structural rules of vehicular protocols.

Phase 2 - Intrusion Detection System (IDS). The second phase is the operational core of the pipeline. Trained on the augmented datasets from the previous phase, the IDS continuously monitors the vehicle's CAN bus for malicious behavior, such as Denial of Service (DoS) attacks, in which malicious nodes flood the network to paralyze critical components like brakes or engines. Given the resource-constrained nature of vehicular edge computing, this module prioritizes lightweight machine learning classifiers capable of low-latency anomaly detection. When an anomaly is detected, the IDS triggers a structured alert containing the raw technical logs.

Phase 3 - Interpretation of Raw Logs. This post-detection phase acts as an interpretative layer within the VSOC. Traditional IDS outputs are mathematical alerts (e.g., a binary flag) accompanied by CAN bus hexadecimal payloads, which are hard for human operators to interpret quickly under pressure. This phase leverages a lightweight LLM configured as an autonomous AI agent. Upon receiving an IDS alert, the agent normalizes raw logs and generates actionable, contextually relevant explanations in natural language, significantly reducing cognitive load and accelerating incidents response time.

3 Methodology, Setup, and Preliminary Results

This section presents the experimental setup (Section 3.1) and the steps followed to evaluate the pipeline combined with their respective preliminary results (Sections 3.2–3.4). *The goal of this feasibility study is to investigate whether current off-the-shelf GenAI models are sufficiently mature to support the proposed architecture.*

3.1 Experimental Setup

To ensure reproducibility and contextualize the computational constraints, the experiments were conducted using the following setup:

- **Dataset:** We used a real-world automotive dataset containing DoS attacks injected via the OBD-II port (the standard physical diagnostic port found under a vehicle's dashboard or steering wheel) of a Hyundai Sonata as an operational baseline, given its high repetition rate and structural predictability. This dataset, provided by the Hacking and Countermeasure Research Lab (HCRL)[2], is a benchmark widely adopted in recent vehicular IDS literature [5] and comprises 3,665,771 messages (3,078,250 normal and 587,521 injected DoS attacks).
- **Infrastructure:** The environment was deployed on Google Colab using the Runtime Version 2026.04, Python 3, and an NVIDIA T4 GPU accelerator.
- **Algorithms & Libraries:** We employed the Pandas library for raw data ingestion and in-memory tabular representation; the Synthetic Data Vault (SDV)[3] library for the CTGAN architecture; Scikit-Learn[4] for the Random Forest classifier; and the Hugging Face transformers, with PyTorch[5], to run the TinyLlama-1.1B-Chat-v1.0[6] model locally.

3.2 Data Synthesis and Semantic Threats

In the first phase, we extracted and tabulated an initial subset of 20,000 rows from the original DoS Attack dataset, comprising the following message types: 14,302 normal (8-byte), 140 normal (2-byte), and 5,558 DoS attacks (8-byte). We configured the training to run for 100 epochs and generated a 100-row synthetic validation sample, which took approximately 3 minutes to process. To evaluate statistical fidelity and whether the generated data match the original distribution, we applied the Kolmogorov-Smirnov test (KS-Test), a metric recommended by Espinosa and Figueira [6]. Across two independent executions, the model achieved Column Shapes Scores up to 77.05% and maintained a stable Column Pair Trends Score of approximately 59%, as detailed in Table 1. Moreover, a visual inspection of the distribution (Figure 2) confirms that while the model successfully captured the dominant presence of the DoS attack (ID '0000'), it also exhibited unexplained variance and over-amplification across several frequency identifiers (e.g., IDs '0131' and '05f0').

However, a manual inspection of the synthetic dataset revealed significant structural limitations, as illustrated in Table 2. As observed in row [05], the categorical traffic flag 'R' appeared in the payload field after the bytes '01 00', with null values (NaNs) filling the remaining fields. Following an audit of the original data,

[2]https://ocslab.hksecurity.net/Datasets/car-hacking-dataset/
[3]https://docs.sdv.dev/
[4]https://scikit-learn.org/
[5]https://pytorch.org/
[6]https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0/

Table 1: CTGAN Statistical Fidelity Metrics (KS-Test)

Metric	Execution 1 (%)	Execution 2 (%)	Variance (%)
Column Shapes	77.05	71.08	~6.0
Column Pair Trends	59.27	58.65	~0.6
Overall Average	68.16	64.87	~3.3

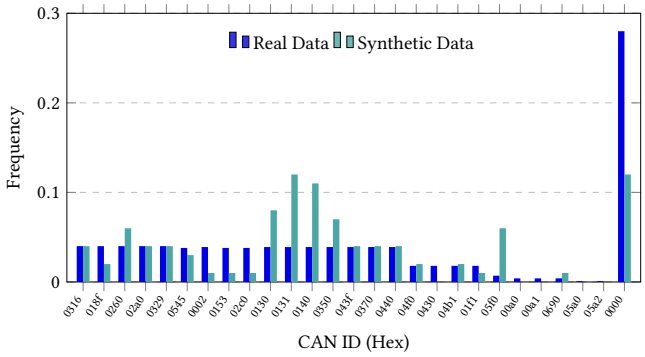

Figure 2: Distribution comparison of 'CAN_ID' between real and CTGAN-generated synthetic data (Representative sample from Execution 1).

we identified that CAN ID 05f0 corresponded to 140 normal short messages (2-byte). Because these messages lacked data for the remaining bytes, the standard tabular loader misaligned the columns, and the CTGAN faithfully reproduced this structure in the generated data. Furthermore, as shown in line [88], the model exhibited semantic violations, incorrectly tagging typical DoS patterns as legitimate traffic.

Table 2: Excerpt of CTGAN Generated Log

Line	CAN ID	DLC	Payload (D0 - D7)	Flag
[01]	0440	8	ff 00 00 00 ff cb 28 4b	R (Normal)
[05]	05f0	2	01 00 **R NaN NaN NaN NaN NaN**	**NaN**
[09]	0000	8	00 00 00 00 00 00 00 00	T (Attack)
[88]	**0000**	8	**00 00 00 00 00 00 00 00**	**R** (Wrong Flag)

3.3 Operational Detection

The second phase validated the operational core (IDS) using a Random Forest classifier (configured with 50 estimators). As observed in Phase 1, the tabular loading of the 50,000-row sample also introduced structural misalignments due to 2-byte messages (383). To prevent these artifacts from biasing the training phase, we excluded these messages. Furthermore, because this subset natively contained a high prevalence of DoS attacks (~23%), the IDS was trained only on real data without requiring synthetic augmentation for this study. The remaining valid dataset (49,617 records) was split into 34,731 instances for training and 14,886 for testing.
As detailed in the confusion matrix (Table 3), the classifier correctly identified all 11,442 normal packets (True Negatives - TN) and all 3,444 attack packets (True Positives - TP). The model produced zero

Table 3: IDS Confusion Matrix (Random Forest on DoS Attack)

	Predicted Normal	Predicted Attack
Actual Normal	11,442 (TN)	0 (FP)
Actual Attack	0 (FN)	3,444 (TP)

False Positives (FP) and zero False Negatives (FN), achieved perfect precision and recall for this specific attack signature within the test sample.

3.4 Raw Logs Interpretation via LLM

In the final phase, we evaluated the feasibility of using TinyLlama as a virtual analyst. The model received alerts triggered by the IDS and was guided by a role-based prompt that included basic threat context. This prompt instructed the model to act as a senior automotive cybersecurity specialist responsible for diagnosis, identifying the threat, explaining its physical risks, and proposing mitigations. We configured the generation with a low temperature (0.3) to favor deterministic responses. Figure 3 presents an example of an IDS alert and its corresponding diagnostic.

Input: Raw IDS Alert (Structured)

```
ALERT_ID: 40522 | PROTOCOL: CAN_BUS
TIMESTAMP: 2016-11-03 18:39:41 UTC
CAN_ID: 0000 | DLC: 8
DATA: 00 00 00 00 00 00 00 00
FLAG: Attack Detected
```

Output: Virtual Analyst Diagnostic (Translated excerpt)

"The alert in the log you provided is a DoS attack signature... The specific type of attack based on the context provided is a spoofing attack... The immediate mitigation action suggested in the context provided is to immediately disable the affected CAN bus. The log also suggests that the vehicle's manufacturer should be contacted..."

Figure 3: Example of IDS Alert (Input) and TinyLlama Output

The text generated by the model exhibited significant operational inconsistencies, including redundancy, logical contradictions, fabricated details, and a lack of fluidity. More critically, the model's explanations were generic and prone to severe hallucinations, generating incorrect risks and mitigation procedures inconsistent with the provided context.

4 Discussion, Conclusion, and Future Work

While the preliminary results confirm the significant potential of GenAI in automotive cybersecurity, they also highlight critical challenges in applying off-the-shelf solutions. Particularly, the structural misalignments during tabular data synthesis suggest that rigorous dataset validation is necessary prior to model training. Additionally, the severe diagnostic hallucinations indicate that AI agents operating without the necessary, explicit automotive context may generate unreliable mitigation strategies. However, these findings should be interpreted within the boundaries of this exploratory

study. The experimental scope was limited to a single attack vector (DoS) and relied on specific baseline configurations, including algorithm choices, fixed sample sizes, and predetermined hyperparameters (e.g., training epochs and LLM temperature), which may not reflect the nuances of real-world vehicular cybersecurity projects.

Consequently, the proposed hybrid architecture serves as a robust conceptual and technical baseline to support future studies evaluating automotive cybersecurity. To mature this framework and overcome the identified constraints, future work should focus on incorporating specific data-driven architectural innovations. Specifically, we propose the following avenues for research:

(1) **Advanced Sanitization and Data Ingestion:** The observed structural anomalies resulted from using a standard tabular loader to handle data loads of varying sizes. Future work should evaluate alternative data parsing libraries and implement robust semantic filters to address potential noise, thereby ensuring clean input data for GenAI training.

(2) **Contextual Grounding:** To mitigate hallucinations and a lack of semantic integrity in interpreting raw data logs, we suggest integrating CAN DBC files and other contextual documents into a Retrieval-Augmented Generation architecture. This would provide broader context and also enable the LLM to accurately translate CAN bus messages. Additionally, exploring advanced prompt engineering techniques and evaluating State-of-the-Art LLMs could improve diagnostic accuracy.

(3) **Federated Learning (FL) and Threat Intelligence:** To ensure the operational scalability of the proposed framework for large-scale vehicle fleets, we suggest that future studies analyze the implementation of a hierarchical FL approach to establish a global, collective threat intelligence network. In this decentralized approach, vehicles collaboratively train local models and share only anonymized weight updates with the VSOC, which, in turn, redistributes the aggregated global detection models to regional nodes and fleets. This bidirectional flow enhances threat intelligence and reduces bandwidth consumption while avoiding the transmission of sensitive raw telemetry data, thereby facilitating compliance with privacy regulations [5].

(4) **IDS with a Defense-in-Depth (DiD):** Rather than testing other baseline classifiers, future architectures can also combine models by applying a DiD strategy. For instance, while Random Forest filters known attacks, Weightless Neural Networks (WNNs), such as the WiSARD model, could act as a secondary layer. Operating on RAM look-up tables, WNNs provide ultra-low hardware consumption and online learning capabilities [12], potentially aiding in detecting unknown attacks that bypass the primary filter.

(5) **End-to-End Streaming Simulation:** Because the current feasibility study evaluated the phases asynchronously, future work should validate the framework as a fully integrated pipeline. Shifting to a dynamic Software-in-the-Loop environment with continuous, real-time vehicular traffic will enable a proper assessment of end-to-end operational latency—from edge-based IDS detection to LLM alert interpretation in a VSOC.

(6) **Expanded Threat Models and Hyperparameter Tuning:** Since this first feasibility study was constrained to a single attack vector (DoS) and fixed configurations, future research must validate the framework against a broader, more diverse set of cyber threats. Furthermore, rigorous hyperparameter optimization should be conducted to balance detection efficacy, computational overhead, and the model's text output.

In conclusion, deploying GenAI in automotive cybersecurity transcends traditional data science; it demands rigorous Software Engineering for AI (SE4AI) practices and robust architectural orchestration to ensure safety and reliability. By addressing the identified semantic and operational gaps, the proposed hybrid framework has the potential to significantly enhance the effectiveness of cyber threat detection systems in vehicular systems and reduce the cognitive load on VSOC analysts.

Artifact Availability

The data and source code, as well as the configurations for the employed models (including training parameters and the TinyLlama prompt), are provided at: **https://zenodo.org/records/21269361**.

References

[1] Rafael Abreu, Emanuel Simão, Carlos Serôdio, Frederico Branco, and Antonio Valente. 2024. Enhancing IoT security in vehicles: A comprehensive review of AI-driven solutions for cyber-threat detection. *AI* (2024). doi:10.3390/ai5040112

[2] R. Aishwarya, V. Vetriselvi, Srinivas Naveen, and Muthuraman A. Ashwin. 2025. An integrated IDS for the Internet of Vehicles using a Large Language Model framework. *Internet of Things* 33 (2025). doi:10.1016/j.iot.2025.101666

[3] MITRE Corporation. 2025. CVE: Common Vulnerabilities and Exposures. https://www.cve.org/. Accessed: Aug. 09, 2026.

[4] João Carmo de Almeida Neto, Leandro Santiago de Araújo, Leopoldo André Dutra Lusquino Filho, and Claudio Miceli de Farias. 2025. Time series forecasting for multidimensional telemetry data based on Generative Adversarial Network in a Digital Twin. *Journal of C. S.* (2025). doi:10.1016/j.jocs.2025.102589

[5] Chloe Soriano de Leon and Cedric Angelo Festin. 2026. Privacy-Preserving Vehicle Intrusion Detection System Using Federated Learning and Homomorphic Encryption. In *Proc. of the Workshop on Computation: Theory and Practice*. Atlantis Press. doi:10.2991/978-94-6239-638-8_31

[6] Erica Espinosa and Alvaro Figueira. 2023. On the Quality of Synthetic Generated Tabular Data. *Mathematics* 11, 15 (2023). doi:10.3390/math11153278

[7] Andre Gheventer, Patricia do Amaral Gurgel, Carlos Henrique Brito, Rafael Maiani de Mello, Sabrina Rocha, Rodrigo Feitosa, and Guilherme Horta Travassos. 2026. Generative AI solutions for software quality: Assessing industrial readiness. *Software Qual J 34, 24* (2026). doi:10.1007/s11219-026-09754-7

[8] Jenny Hofbauer, Kevin Klaus Gomez Buquerin, and Hans-Joachim Hof. 2023. From SOC to VSOC: Transferring key requirements for efficient vehicle security operations. In *Proc. of the 21st escar Europe (2023)*. doi:10.13154/294-10389

[9] ISO/SAE. 2021. ISO/SAE 21434:2021. Road vehicles — Cybersecurity engineering.

[10] Aristeidis Karras, Leonidas Theodorakopoulos, Christos Karras, and Alexandra Theodoropoulou. 2026. Towards LLM-Driven Cybersecurity in Autonomous Vehicles: A Big Data-Empowered Framework with Emerging Technologies. *Machine Learning and Knowledge Extraction* 8, 2 (2026), 43. doi:10.3390/make8020043

[11] Junhui Li, Nikolaos Ersotelos, Michail-Antisthenis Tsompanas, and Gregory Epiphaniou. 2026. Challenges and opportunities of synthetic data generation for machine learning-based intrusion detection systems in in-vehicle networks. *Vehicular Communications* 58 (2026), 100998. doi:10.1016/j.vehcom.2025.100998

[12] Zachary Susskind, Aman Arora, Alan T. L. Bacellar, Diego L. C. Dutra, Igor D. S. Miranda, Mauricio Breternitz, Priscila M. V. Lima, Felipe M. G. França, and Lizy K. John. 2023. An FPGA-Based Weightless Neural Network for Edge Network Intrusion Detection. In *Proc. of the 2023 ACM/SIGDA Int. Symp. on Field Programmable Gate Arrays (FPGA '23)*. ACM. doi:10.1145/3543622.3573140

[13] UNECE. 2021. UN Regulation No. 155 - Cyber security and cyber security management system. United Nations. E/ECE/TRANS/505/Rev.3/Add.154.

[14] Xiang Wang, Jian Zhao, Pengbo Liu, Nianmin Yao, and Zheng Xu. 2026. CAN Bus Intrusion Detection Based on Deep Learning With Data Augmentation for Connected Autonomous Vehicles. *IEEE Transactions on Vehicular Technology* (2026). doi:10.1109/TVT.2025.3603056