Intrusion Detection System in CAN Messages

Project Report Submitted
in partial fulfilment of the requirements for the ward of the

Bachelor of Technology

by

Abhimanyu Tripathi (B222003)

Anshuman Mahabhoi (B422010)

Saumyajeet Varma (B522053)

Under the supervision of

Dr. Puspanjali Mohapatra

Department of Computer Science and Engineering
International Institute of Information Technology, Bhubaneswar

APPROVAL OF THE VIVA-VOCE BOARD

Thu, 20 November 2025

Certified that the report entitled **"Intrusion Detection System in CAN Messages"** submitted by **Abhimanyu Tripathi (B22003)**, **Anshuman Mahabhoi (B422010)** and **Saumyajeet Varma (B522053)** to **International Institute of Information Technology, Bhubaneswar** in partial fulfillment of the requirements for **Technical Writing in Computer Science Engineering (7th Semester)** under the BTech Programme has been accepted by the examiners during the viva-voce examination held today.

(Supervisor) (Panel Head)

(Internal Examiner 1) (Internal Examiner 2)

CERTIFICATE

This is to certify that the report entitled **"Intrusion Detection System in CAN Messages"** submitted by **Abhimanyu Tripathi (B222003)**, **Anshuman Mahabhoi (B422010)** and **Saumyajeet Varma (B522053)** to **International Institute of Information Technology, Bhubaneswar** is a record of bonafide project work under my supervision, and the report is submitted for end-semester evaluation of **Technical Writing, B.Tech, 7th Semester.**

Dr. Puspanjali Mohapatra
(Supervisor)

DECLARATION

We certify that

1. The work contained in the report has been done by me under the general supervision of my supervisor.

2. The work has not been submitted to any other Institute for any degree or diploma.

3. I have followed the guidelines provided by the Institute in writing the thesis.

4. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

5. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

6. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Abhimanyu Tripathi	Anshuman Mahabhoi	Saumyajeet Varma
(B222003)	(B422010)	(B522053)

ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to all those who have guided us throughout the development of this report. This work would not have been possible without the continuous guidance, invaluable insights, and dedicated support of our esteemed mentor, **Dr. Puspanjali Mohapatra**, who supervised us closely during the project.
We would also like to extend our sincere appreciation to our institution for providing the encouragement and resources necessary for this undertaking. This experience has allowed us to gain significant knowledge and exposure in our field, enriching our understanding and enhancing our skills.

Abhimanyu Tripathi
(B22003)

Anshuman Mahabhoi
(B422010)

Saumyajeet Varma
(B522053)

Contents

1 Introduction

The rapid integration of electronic control units (ECUs) in modern vehicles has fundamentally transformed the automotive landscape, making contemporary cars complex cyber-physical systems. These ECUs are interconnected through the Controller Area Network (CAN), a lightweight and efficient communication protocol originally designed to ensure fast, reliable, and real-time data exchange between vehicle subsystems. However, despite its technical strengths, CAN was developed at a time when vehicular connectivity was minimal, and cybersecurity threats were largely nonexistent. As a result, the protocol lacks essential security mechanisms such as message authentication, encryption, and source verification. With the emergence of connected, autonomous, and software-defined vehicles, this absence of intrinsic security poses a significant challenge. Newer vehicles now integrate internet connectivity, telematics systems, Wi-Fi, Bluetooth, and Vehicle-to-Everything (V2X) communication, which collectively broaden the attack surface available to cyber adversaries. Thus, malicious individuals can exploit these external interfaces to infiltrate the vehicle's internal network and inject unauthorized CAN messages that may influence critical functions including steering, braking, and engine control. The increasing frequency and sophistication of such attacks emphasize the urgent need for robust intrusion detection systems (IDS) specifically tailored for automotive networks.

1.1 Need for Automotive Intrusion Detection

The vulnerability of the CAN protocol stems primarily from its lack of built-in security features. Since the protocol does not authenticate message sources, any compromised ECU or external device with access to the network can transmit arbitrary messages posing as legitimate components. This allows attackers to introduce falsified engine speed readings, spoof gear positions, override sensor values, or flood the bus to suppress legitimate communication, thereby creating potentially dangerous driving scenarios. Additionally, the widespread adoption of internet-enabled infotainment systems and wireless communication modules has resulted in vehicles being constantly exposed to external networks. Every wireless interface—from telematics units to keyless entry systems—introduces a possible entry point for cyberattacks. Unlike traditional computing environments, vehicular systems require strict real-time operation and cannot afford delays, making it impractical to rely on heavy cryptographic operations or conventional IT security techniques. The safety-critical nature of automotive systems further heightens the importance of continuous monitoring of CAN traffic to identify anomalies indicative of cyberattacks. Regulatory frameworks such as ISO/SAE 21434 and UNECE WP.29 have recognized these risks and now mandate that manufacturers implement vehicle cybersecurity measures, including detection and mitigation of malicious intrusions. These factors collectively underline the necessity for dedicated IDS solutions capable of analyzing CAN-bus

behavior and ensuring the integrity of communication within the vehicle.

1.2 Scope of the Study

The scope of this introductory study is focused on examining the cybersecurity vulnerabilities of CAN-bus networks and understanding the motivation, challenges, and existing research approaches related to in-vehicle intrusion detection. This report does not propose a new detection model; instead, it serves as a foundational analysis of the problem domain. It outlines why automotive IDS solutions are essential, the limitations of the CAN protocol, and the specific characteristics of cyberattacks that threaten in-vehicle systems. Additionally, the study explores the evolution of in-vehicle networking and discusses how increasing vehicle connectivity has simultaneously increased exposure to cyber threats. The literature reviewed in this report spans reinforcement learning-based IDS models, generative models for data augmentation, adversarial multi-agent IDS frameworks, and deep learning techniques tailored for CAN-bus environments. By summarizing existing research and identifying common gaps, the report establishes a strong contextual basis for further exploration of advanced IDS methodologies for securing modern vehicles.

1.3 Evolution of In-Vehicle Network Security

Automotive communication systems have undergone substantial evolution over the past three decades, each phase bringing enhanced functionality but also new cybersecurity concerns. In the early years, vehicles contained only a handful of ECUs, typically operating in isolation and responsible for basic mechanical functions. During this era, external connectivity was almost nonexistent, and the risk of cyber intrusion was negligible. As vehicle functionality grew more complex, manufacturers introduced the CAN bus to interconnect ECUs responsible for powertrain, braking, transmission, and comfort systems. Although CAN efficiently supported real-time communication, its designers did not incorporate security mechanisms because the system was intended to be closed and physically inaccessible to attackers. The next phase of evolution introduced connected car technologies. Infotainment systems capable of internet access, telematics units, GPS navigation, and Bluetooth interfaces created communication pathways that extended beyond the physical boundary of the vehicle. Academic and industry researchers soon demonstrated that these external interfaces could be exploited to gain remote access to the internal CAN bus. The widely publicized Jeep Cherokee hack exemplified how an attacker could manipulate critical driving functions remotely by exploiting vulnerabilities in the vehicle's wireless systems. The move toward software-defined and autonomous vehicles further amplified these challenges. Autonomous systems depend on a massive network of sensors, cameras, LiDAR, and high-performance ECUs that must communicate seamlessly to ensure safe decision-making. This dense digital ecosystem creates

numerous vectors for cyberattacks targeting both perception and control mechanisms. Consequently, the evolution of in-vehicle networks has revealed an urgent need for sophisticated intrusion detection approaches capable of monitoring CAN traffic, detecting deviations from normal patterns, and responding to malicious activity in real time.

2 Literature Survey

2.1 MAFSIDS: A Reinforcement Learning-Based Intrusion Detection Model

2.1.1 Overview

MAFSIDS [1]is an intrusion detection framework that combines graph-based deep learning with reinforcement learning to achieve efficient feature selection and high classification performance on large-scale intrusion datasets. The central idea of this work is to reduce redundant features through intelligent selection while maintaining strong detection accuracy across diverse network attack types.

2.1.2 Methodology

The approach begins with the use of a Graph Convolutional Network (GCN) to understand relational dependencies among dataset features. This graph-based embedding helps identify correlations and provides a structured representation of the data. Feature selection is performed through a Multi-Agent Feature Selection (MAFS) framework, where reinforcement learning agents collaboratively determine which features should be retained or discarded. The final classification stage employs a Deep Q-Network (DQN) trained on the optimized feature set to categorize traffic instances as normal or malicious.

2.1.3 Results

The model demonstrates strong performance on both CSE-CIC-IDS2018 and NSL-KDD datasets, achieving approximately 96.8% and 99.1% accuracy respectively. The reinforcement-learning-driven feature selection mechanism reduces redundant attributes by nearly 80%, leading to significant computational savings without degrading detection accuracy.

2.1.4 Strengths

The principal advantage of MAFSIDS lies in its combination of structural feature learning and dynamic RL-based optimization, enabling effective dimensionality reduction and improving system efficiency. The use of GCN allows the system to capture complex inter-feature relationships that traditional feature selection methods often miss.

2.1.5 Limitations

Despite its strengths, the model is computationally expensive due to the involvement of multiple RL agents and GCN training. Its reliance on predefined reward signals makes it sensitive to design choices, and the framework may require extensive tuning when applied to datasets with different feature distributions.

2.2 ID-RDRL: Recursive Deep Reinforcement Learning for Intrusion Detection

2.2.1 Overview

ID-RDRL [2]aims to enhance intrusion detection by framing feature selection as a learning problem, enabling the system to recursively identify the most informative subset of attributes. This is intended to improve classification quality while reducing computational overhead in handling large feature spaces.

2.2.2 Methodology

The system integrates Recursive Feature Elimination (RFE) with a Deep Q-Network (DQN). RFE provides an initial estimate of feature importance, while the DQN agent treats each possible selection as a state and learns through trial-and-error which features should be retained. The agent receives reward signals based on classification accuracy and iteratively converges toward an optimal set of features. A final classifier is trained on the refined dataset to detect different types of attacks.

2.2.3 Results

Experiments conducted on NSL-KDD and UNSW-NB15 datasets show high accuracy, with values reaching 98.2% and 96.5% respectively. F1-scores also remain consistently high, demonstrating strong performance across both balanced and imbalanced datasets.

2.2.4 Strengths

The integration of RL enables ID-RDRL to discover non-linear feature interactions and adapt automatically to the dataset's structure. Its recursive elimination strategy ensures robustness and prevents over-reliance on noisy or redundant features.

2.2.5 Limitations

A key limitation lies in the model's high training cost, as multiple evaluations are required during the RL-driven elimination process. Additionally, the final performance heavily depends on the reward formulation and the base learner used in the RFE.

2.3 Deep Reinforcement Learning IDS Using CIC-IDS2017

2.3.1 Overview

This work [3] employs deep reinforcement learning to build an adaptive intrusion detection system capable of handling dynamic and evolving network threats. The study uses the CIC-IDS2017 dataset, which offers realistic and diverse network traffic patterns.

2.3.2 Methodology

The model encodes network flow characteristics into state representations for a Deep Q-Network (DQN). The agent learns classification policies through reward-driven optimization, where correct classifications earn positive rewards and incorrect decisions produce penalties. Over time, the system refines its internal decision boundaries without the need for explicit retraining.

2.3.3 Results

The DRL-based system achieves approximately 94.5% accuracy and maintains an F1-score of around 93%. These results suggest that reinforcement learning can adapt well to shifting traffic patterns and detect several attack categories, including DDoS and brute force attacks.

2.3.4 Strengths

The primary strength of this model is its ability to adjust dynamically to changes in traffic behavior, making it suitable for real-time environments. Reinforcement learning reduces the need for frequent retraining, which is beneficial in large-scale networks..

2.3.5 Limitations

A significant limitation is that RL models can exhibit unstable learning patterns if the reward structure is not carefully designed. In addition, performance relies heavily on effective state representation, which may be challenging to construct for complex traffic scenarios.

2.4 XGBoost-Based Feature Selection for UNSW-NB15

2.4.1 Overview

This study [9] investigates how feature selection based on XGBoost can improve the efficiency and performance of traditional machine learning models on the UNSW-NB15 dataset.

2.4.2 Methodology

XGBoost is used to compute feature importance scores, which serve as criteria for pruning low-value features. After reducing dimensionality, several classifiers—including decision trees, ANN, logistic regression, SVM, and kNN—are trained on the optimized dataset and evaluated on both binary and multiclass tasks.

2.4.3 Results

The study reports improved performance for certain models, most notably a rise in decision tree accuracy from approximately 88% to over 90%. Feature selection also reduces computation time.

2.4.4 Strengths

The method is simple yet effective, demonstrating that feature-selection can enhance accuracy and efficiency, especially in resource-limited environments.

2.4.5 Limitations

Reliance on XGBoost's importance scoring may lead to the removal of subtle features that contribute to non-linear patterns. The approach may not generalize equally well across all classifiers.

2.5 Zero-Day Detection Using Autoencoder-Based Hybrid Models

2.5.1 Overview

This research [5]focuses on identifying unknown attacks using anomaly detection techniques that rely on autoencoder reconstruction behavior, combined with supervised learning classifiers

2.5.2 Methodology

An autoencoder is trained exclusively on benign data to learn its underlying structure. When anomalous inputs are processed, their reconstruction error is significantly higher. These errors are then used as input features for Random Forest and XGBoost models, forming hybrid classifiers capable of detecting zero-day attacks.

2.5.3 Results

The model demonstrates exceptional accuracy on the CIC-MalMem-2022 dataset, with Random Forest-AE achieving 99.9892% accuracy and XGBoost-AE achieving 99.9741% accuracy.

2.5.4 Strengths

The approach excels at detecting previously unseen threats because it relies on anomaly-based modeling instead of signature-based classification

2.5.5 Limitations

Autoencoders can inadvertently reconstruct certain malicious patterns too well, reducing anomaly sensitivity. Performance also depends on how clean the benign training data is.

2.6 Attention-CNN-LSTM IDS for In-Vehicle Networks

2.6.1 Overview

This study [7] develops an IDS optimized for automotive environments by integrating convolutional networks, recurrent sequence modeling, and attention mechanisms.

2.6.2 Methodology

CAN traffic is transformed into temporal sequences. CNN layers extract spatial features from each timestep, LSTM layers model sequential dependencies, and the attention mechanism highlights the most relevant segments contributing to anomalies. The model is trained using labeled in-vehicle datasets.

2.6.3 Results

The system achieves high performance, including 99.43% accuracy and near-perfect ROC-AUC scores, demonstrating suitability for real-time automotive IDS applications.

2.6.4 Strengths

Its hybrid architecture effectively captures both local and temporal patterns, while attention improves interpretability and precision.

2.6.5 Limitations

The model demands considerable computational power and large labeled datasets, which may limit deployment on embedded systems.

2.7 Cognitive-Based IDS Using DNN + SVM

2.7.1 Overview

This research [8] proposes a hybrid IDS that combines deep learning-based feature extraction with SVM-based classification to enhance the precision of anomaly detection.

2.7.2 Methodology

The model preprocesses data from the KDD99 dataset, using a deep neural network (DNN) to learn hierarchical features that are then fed into an SVM classifier. The DNN extracts high-level representations while the SVM performs the final decision-making..

2.7.3 Results

The system achieves a classification accuracy of 95.4%, with stable performance across different attack categories.

2.7.4 Strengths

The integration of DNN and SVM enhances detection stability and allows the model to capture non-linear feature interactions while maintaining strong decision boundaries.

2.7.5 Limitations

Its performance on modern datasets remains uncertain due to the aging nature of KDD99. The dual-model pipeline may also increase computation overhead.

3 Comparitive Study

Table 1: Comparison of Research Papers on Intrusion Detection Systems

S.No	Paper Title	Dataset Used	Methodology / Models Used	Novelty / Contribution	Accuracy / Results
1	MAFSIDS: Reinforcement Learning-Based IDS	CSE-CIC-IDS2018, NSL-KDD	GCN + Multi-Agent Feature Selection + DQN	80% feature reduction; efficient hybrid IDS	96.8%, 99.1%
2	ID-RDRL: Recursive Deep Reinforcement Learning IDS	NSL-KDD, UNSW-NB15	RFE + DQN	Dynamic RL-driven feature selection	98.2%, 96.5%
3	Deep Reinforcement Learning IDS	CIC-IDS2017	Deep Q-Network (DQN)	Adaptive real-time learning	94.5% (Acc), F193%
4	XGBoost-Based Feature Selection IDS	UNSW-NB15	XGBoost feature ranking + ML models	Feature reduction improved performance	90.85% (Binary, DT)
5	Zero-Day Detection Hybrid AE Models	CIC-MalMem-2022	Autoencoder + RF / XGBoost	Zero-day anomaly detection	99.98%, 99.97%
6	ACL-IDS for In-Vehicle Networks	CIC-MalMem-2022	CNN + LSTM + Attention	Lightweight real-time IDS	99.43%
7	Cognitive-Based IDS	KDD99	DNN + SVM	Improved precision and model stability	95.4%

4 Motivation

4.1 Growth of Intelligent Transportation Systems

Modern vehicles are rapidly transitioning into intelligent and automated systems equipped with advanced Electronic Control Units (ECUs), sensors, and communication modules. This transformation has increased dependency on in-vehicle networks such as the Controller Area Network (CAN) bus for real-time data exchange and cooperative decision-making. As vehicles continue to evolve into connected and autonomous platforms, ensuring secure internal communication has become a critical necessity to safeguard passenger safety and system reliability.

4.2 Vulnerabilities in the CAN Communication Protocol

The CAN bus protocol, despite being the backbone of intra-vehicle communication, lacks essential cybersecurity features such as message authentication and encryption. It operates on a broadcast mechanism where all ECUs receive messages without verifying their source. Consequently, attackers can easily inject malicious messages, modify existing data, or flood the network, triggering safety-critical malfunctions like disabling brakes, altering RPM readings, or manipulating gear signals. These vulnerabilities highlight the urgent need for intelligent and robust anomaly detection approaches.

4.3 Limitations of Existing Intrusion Detection Approaches

Traditional Intrusion Detection Systems (IDS) based on rule-based techniques, signature matching, and statistical analysis fail to detect new or evolving attack patterns. Similarly, supervised machine learning methods require large labeled attack datasets, which are costly, time-consuming to collect, and often unavailable. Many existing models struggle with:

- High false-positive rates due to complex CAN traffic dynamics
- Poor adaptability to unseen attacks
- Manual feature engineering requirements
- Inability to operate effectively in real-time environments

These limitations strongly motivate the need for an adaptive and scalable detection framework.

4.4 Need for Advanced AI-Based Intrusion Detection

Deep learning-based anomaly detection has shown significant potential in identifying subtle deviations in CAN traffic. However, models trained only on normal or limited attack data lack generalization capabilities. Generative Adversarial Networks (GANs) enable learning normal traffic distributions and generating synthetic attack samples, whereas Reinforcement Learning can optimize feature selection automatically, enhancing model performance. A hybrid approach combining RL and GAN can:

- Learn complex internal structures of CAN data
- Detect previously unseen attacks efficiently
- Reduce dependency on large labeled datasets
- Improve robustness and detection accuracy

5 Objectives

5.1 To Strengthen Security in Modern Intelligent Vehicles

As automotive systems evolve toward connected and autonomous architectures, ensuring secure in-vehicle communication has become essential for passenger safety and operational reliability. A key objective of this research is to support the development of secure transportation ecosystems by enabling early detection of malicious behaviors that may compromise critical vehicular control systems.

5.2 To Address Vulnerabilities in CAN Bus Communication

The CAN bus protocol, despite its efficiency and reliability, lacks built-in cybersecurity mechanisms such as encryption and message authentication. This exposes the communication network to cyber-attacks including message injection, spoofing, and DoS attacks. Therefore, an important objective is to investigate existing weaknesses in CAN communication and propose effective mechanisms for detecting abnormal traffic patterns that could pose safety risks.

5.3 To Improve the Effectiveness of Intrusion Detection Systems

Traditional IDS approaches often struggle to detect evolving attack types and produce high false alarm rates due to the dynamic nature of CAN traffic. This research aims to develop a more accurate and adaptive intrusion detection strategy capable of identifying known and unknown attack patterns with improved precision and reliability, overcoming limitations of existing rule-based and statistical detection methods.

5.4 To Utilize Data-Driven Techniques for Enhanced Anomaly Detection

With the increasing availability of vehicle CAN datasets, machine learning and deep learning techniques provide new opportunities for intelligent cybersecurity solutions. This research seeks to explore advanced data-driven methodologies for analyzing CAN traffic behavior and identifying anomalies by modeling real-world communication patterns rather than relying solely on handcrafted rules.

5.5 To Facilitate Real-Time and Scalable Implementation

An additional objective is to ensure that the proposed detection method is computationally efficient and capable of operating in real time without interrupting vehicular communication. The research aims to design a scalable and practical intrusion detection solution suitable for deployment in real automotive environments and adaptable to various vehicle models and attack scenarios.

5.6 To Contribute to Future Advancement of Automotive Cybersecurity

Finally, this research intends to generate meaningful insights and contribute to ongoing development efforts in automotive security. By evaluating the effectiveness of advanced

anomaly detection methods on real datasets, the study aims to provide guidance for future improvements and encourage further innovation in the protection of intelligent transportation systems.

6 Methodology

The methodology adopted in this study consists of two major components: the construction and analysis of CAN-bus datasets derived from real vehicular environments, and the development of a hybrid anomaly detection framework that integrates Reinforcement Learning (RL) with a Generative Adversarial Network (GAN). This section presents a detailed explanation of the datasets, preprocessing operations, model architecture, training strategy, and testing procedures that together form the complete methodological pipeline.

6.1 DATASET

We used car-hacking datasets which include DoS attack, fuzzy attack, spoofing the drive gear, and spoofing the RPM gauge. Datasets were constructed by logging CAN traffic via the OBD-II port from a real vehicle while message injection attacks were performing. Datasets contain each 300 intrusions of message injection. Each intrusion performed for 3 to 5 seconds, and each dataset has total 30 to 40 minutes of the CAN traffic.

- DoS Attack : Injecting messages of '0000' CAN ID every 0.3 milliseconds. '0000' is the most dominant.
- Fuzzy Attack : Injecting messages of totally random CAN ID and DATA values every 0.5 milliseconds.
- Spoofing Attack (RPM/gear) : Injecting messages of certain CAN ID related to RPM/gear information every 1 millisecond.

6.1.1 Data attributes

Timestamp, CAN ID, DLC, DATA[0], DATA[1], DATA[2], DATA[3], DATA[4], DATA[5], DATA[6], DATA[7], Flag

- Timestamp : recorded time (s)
- CAN ID : identifier of CAN message in HEX (ex. 043f)
- DLC : number of data bytes, from 0 to 8
- DATA[0 7] : data value (byte)
- Flag : T or R, T represents injected message while R represents normal message

Table 2: CAN dataset statistics (per attack trace): total messages, normal messages, and injected messages

Attack Type / Dataset	**of Messages**	**of Normal Messages**	**of Injected Messages**
DoS Attack	3,665,771	3,078,250	587,521
Fuzzy Attack	3,838,860	3,347,013	491,847
Spoofing the drive gear	4,443,142	3,845,890	597,252
Spoofing the RPM gauge	4,621,702	3,966,805	654,897
Attack-free (normal)	988,987	988,872	–

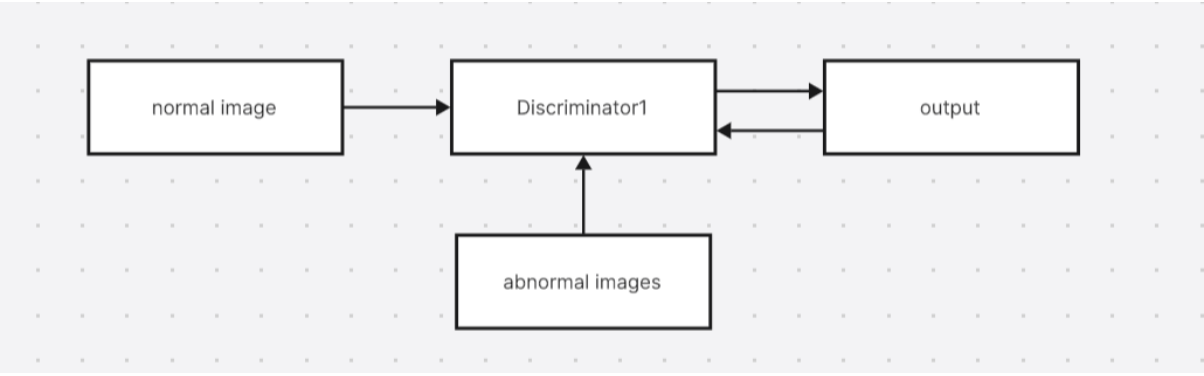
6.1.2 Data Preprocessing

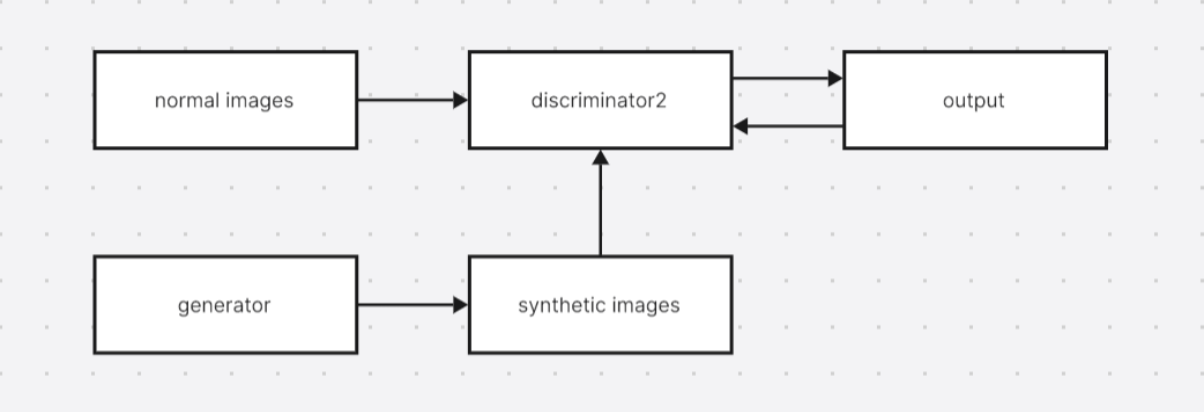
Preprocessing of the CAN traffic begins with parsing each log entry into structured representations. CAN IDs and data bytes are converted from hexadecimal to integer values to facilitate numerical learning operations. Messages are then encoded into machine-interpretable feature vectors using one-hot encoding, particularly for CAN IDs, which contain categorical information. To make CAN IDs compatible with image-based representation learning, each hexadecimal character of the ID is one-hot encoded and arranged into a two-dimensional 16×3 matrix, effectively converting each CAN identifier into a compact image-like representation.

The resulting matrices are resized to fixed dimensions to match the expected input shape of the neural network models. This resizing is performed in batch using a shape-standardization function. The complete preprocessed dataset therefore consists of normalized, vectorized, and image-structured CAN messages ready for adversarial training and classification.

6.2 Model Architecture

The detection framework employed in this study integrates Reinforcement Learning (RL) with a Generative Adversarial Network (GAN) to produce an adaptive anomaly detection system capable of modeling normal CAN traffic distribution and distinguishing malicious patterns. The model consists of three principal components: the generator, the discriminators, and an RL-based optimization loop.





6.2.1 Generator Network

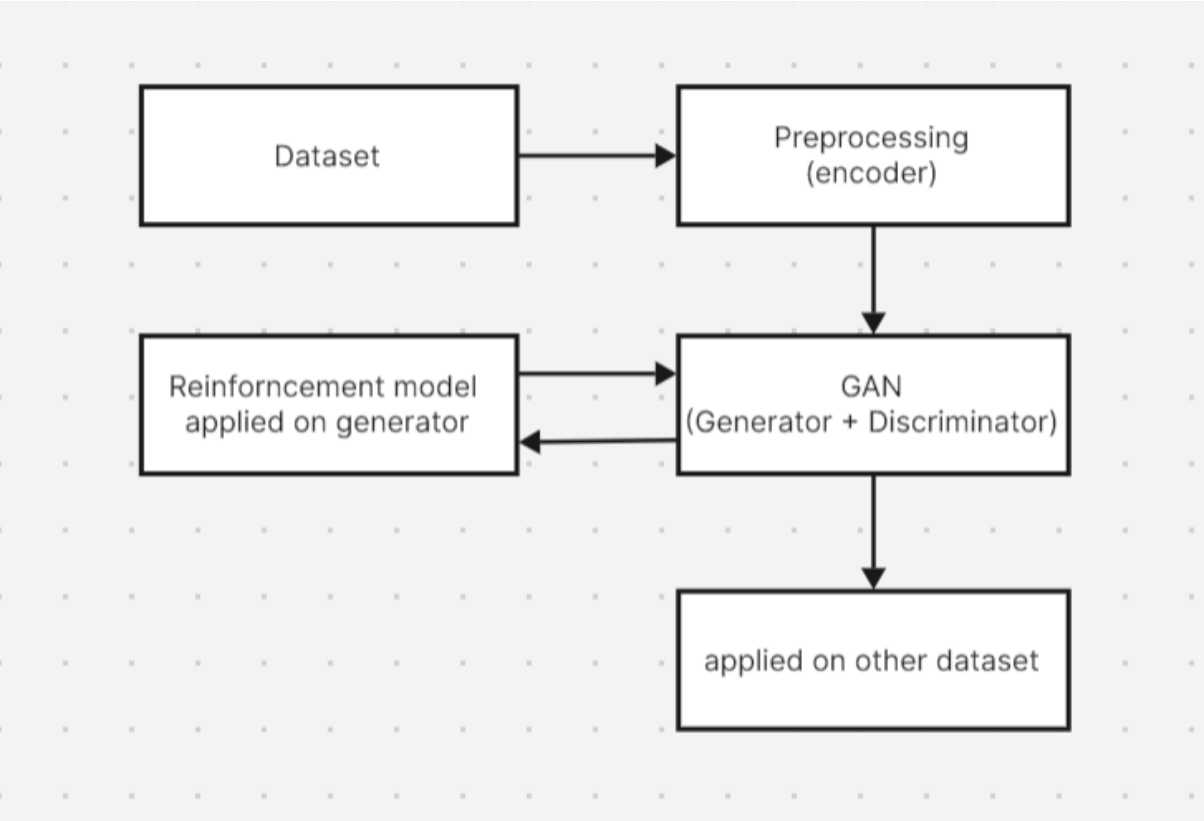
The generator is a fully connected neural network designed to create synthetic CAN ID images that mimic the distribution of real, benign CAN traffic. The input to the generator is a latent noise vector sampled from a multivariate normal distribution. The generator employs stacked dense layers with ReLU activation, followed by batch normalization, to stabilize training and ensure smooth gradient flow. The output of the generator is a synthetic CAN image with the same 16×3 structure used for the real CAN ID images.

6.2.2 Discriminator Networks

Two discriminator models are employed. The first discriminator is responsible for differentiating real CAN ID images from synthetic images generated during training. It uses a convolutional or dense architecture with binary cross-entropy loss to classify images as real or fake. The second discriminator is trained specifically to classify attack images and to detect unknown or evolving attack patterns by evaluating synthetic images produced by the generator. This two-stage discriminator design enables the model to learn fine-grained distinctions between normal and anomalous patterns as well as improve the generator's ability to produce realistic variants of attack inputs.

6.2.3 Reinforcement Learning Integration

Reinforcement Learning is incorporated to enhance the generator's training process by providing reward-driven optimization. Instead of updating the generator solely through backpropagation from the discriminator's gradients, the model computes rewards based on the discriminator's output for synthetic samples. These rewards represent how convincingly the generator is able to imitate real CAN patterns. The generator uses policy-gradient-style updates, where the reward signal encourages the production of samples that move closer to the true distribution of benign CAN ID images. This RL-enhanced feedback loop improves the generator's ability to model complex, non-linear CAN dynamics and leads to more robust synthetic sample generation, which in turn strengthens the discriminator's detection ability.



6.3 Training Procedure

The training process alternates between discriminator updates and generator updates. First, the discriminator is trained with both real normal images and attack images to differentiate between the two categories. It is also trained on synthetic samples produced by the generator to classify them as fake. By training simultaneously on real attacks, normal messages and synthetic anomalies, the discriminator learns a rich decision boundary that separates normal traffic from malicious patterns.

During generator training, latent noise vectors are passed through the generator

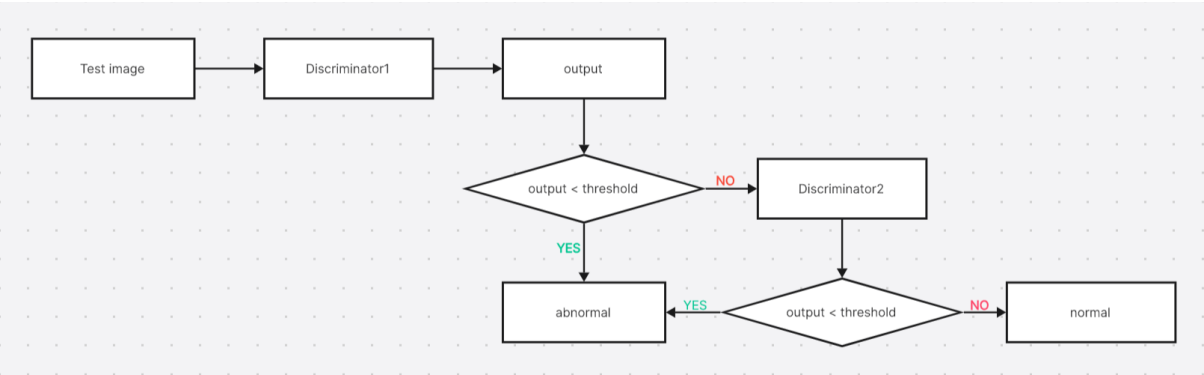
to create synthetic samples. These samples are evaluated by the discriminator, and a reward is computed based on the discriminator's probability output. The generator parameters are updated using reinforcement learning techniques that attempt to maximize the cumulative reward, effectively encouraging the generator to produce more realistic CAN traffic representations.

This adversarial and reinforcement learning process continues iteratively, gradually improving both the generator's realism and the discriminator's discriminative strength. The final model parameters are obtained after convergence of both networks.

6.4 Testing and Evaluation

During testing, only the discriminators are used. Incoming CAN messages are preprocessed using the same pipeline as the training data and fed into the discriminator network responsible for anomaly detection. The output is a classification label indicating whether the message is normal or indicative of an attack. Evaluation metrics such as accuracy, precision, recall, and F1-score are computed using ground truth labels. Zero-day detection capability is assessed by withholding certain attack categories during training and evaluating model performance on unseen attack patterns.

The complete methodology thus provides a structured approach for detecting CAN-bus anomalies by combining GAN-based sample generation, reinforcement learning-based optimization and supervised anomaly discrimination. This hybrid design allows the system to generalize better to unseen attack patterns and improves detection robustness in highly dynamic vehicular network environments.



7 Conclusion

The increasing integration of advanced electronics, automation, and connectivity in modern vehicles has transformed the automotive industry, but it has also introduced significant cybersecurity challenges. The vulnerability of the CAN bus communication protocol to various forms of cyber-attacks highlights the urgent need for effective and intelligent intrusion detection mechanisms to ensure the safety, reliability, and resilience of vehicular

systems. Existing traditional detection techniques, though valuable, often fail to provide sufficient accuracy, adaptability, and scalability required to combat newly emerging and sophisticated attack patterns.

This research emphasizes the importance of adopting advanced data-driven approaches capable of understanding complex CAN traffic behavior and distinguishing legitimate communication from malicious anomalies. Through exploring real vehicular datasets, analyzing attack characteristics, and evaluating detection strategies, the study contributes toward building a more secure and future-ready automotive environment. Strengthening CAN bus security not only protects critical vehicle functions but also builds trust in intelligent transportation systems, supporting the safe deployment of connected and autonomous vehicles.

In summary, continuous research, development, and implementation of intelligent intrusion detection technologies are essential to enhance automotive cyber defense. Future work may extend this study by exploring more robust datasets, integrating hardware-based security enhancements, and designing real-time deployment frameworks that support next-generation vehicle networks. A collaborative effort between researchers, automotive manufacturers, and cybersecurity experts will be crucial in advancing secure, reliable, and resilient transportation infrastructures.

References

[1] K. Ren, Z. Li, X. Wang and Y. Zhang, "MAFSIDS: A reinforcement learning-based intrusion detection model for multi-intelligence feature selection networks," *Journal of Big Data*, vol. 10, Article 98, 2023.

[2] K. Ren, J. Shen, X. Zhao, et al., "ID-RDRL: A deep reinforcement learning-based feature selection intrusion detection model," *Scientific Reports*, vol. 12, Article number (2022).

[3] C. Strickland, C. Saha, M. Zakar, S. Nejad, N. Tasnim, D. Lizotte and A. Haque, "DRL-GAN: A hybrid approach for binary and multiclass network intrusion detection,"

[4] M. Mouyart, G. Medeiros Machado and J.-Y. Jun, "A Multi-Agent Intrusion Detection System Optimized by a Deep Reinforcement Learning Approach with a Dataset Enlarged Using a Generative Model to Reduce the Bias Effect," *Journal of Sensor and Actuator Networks*, vol. 12, no. 5, 2023. DOI: 10.3390/jsan12050068.

[5] Z. Dai, Y. Yang, H. Zhang, et al., "An intrusion detection model to detect zero-day attacks in unseen data using machine learning," *(open access)*, 2024 — autoencoder + supervised hybrid methodology evaluated on CIC-MalMem-2022. A

[6] W. Yang, A. Acuto, Y. Zhou and D. Wojtczak, "A Survey for Deep Reinforcement Learning Based Network Intrusion Detection," arXiv:2410.07612, Sep. 2024.

[7] A. M. Alashjaee, et al., "Attention-CNN-LSTM based intrusion detection (ACL-IDS): hybrid deep model for network/vehicle traffic," *Scientific Reports / related venue*, 2025. (See published work describing CNN+LSTM+attention for network/vehicle traffic; referenced in surveys).

[8] S. Parhizkari, "A cognitive-based method for intrusion detection systems: Deep neural features with SVM classification," arXiv:2005.09436, 2020.

[9] S. M. Kasongo, "Performance Analysis of Intrusion Detection Systems Using Machine Learning Techniques and Feature Selection on UNSW-NB15," Doctoral thesis / technical report, 2020.